AN EFFICIENT INDEXING AND RETRIEVAL OF IRIS BIOMETRICS DATA USING HYBRID TRANSFORM AND FIREFLY BASED K-MEANS ALGORITHM TITLE

EMAD TAHA KHALAF

Doctor of Philosophy (Computer Science)

UNIVERSITI MALAYSIA PAHANG

UNIVERSITI MALAYSIA PAHANG

NOTE : * If the thesis is CONFIDENTIAL or RESTRICTED, please attach a thesis declaration letter.

SUPERVISOR’S DECLARATION

I hereby declare that I have checked this thesis and in my opinion, this thesis is adequate in terms of scope and quality for the award of the degree of Doctor of Philosophy (Computer Science)

____

(Supervisor’s Signature)

Full Name : DR.MUAMER N. MOHAMMED

Position : SENIOR LECTURER

Date :

STUDENT'S DECLARATION

I hereby declare that the work in this thesis is based on my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously or concurrently submitted for any other degree at Universiti Malaysia Pahang or any other institutions.

____

(Student's Signature)

Full Name : EMAD TAHA KHALAF

ID Number : PCC13010

Date :

AN EFFICIENT INDEXING AND RETRIEVAL OF IRIS BIOMETRICS DATA USING HYBRID TRANSFORM AND FIREFLY-BASED K-MEANS ALGORITHM

EMAD TAHA KHALAF

Thesis submitted in fulfillment of the requirements
for the award of the degree of
Doctor of Philosophy
(Computer Science)

Faculty of Computer Systems and Software Engineering

UNIVERSITI MALAYSIA PAHANG

MARCH 2019

ACKNOWLEDGEMENTS

I am immeasurably indebted to ALMIGHTY ALLAH, the propitious, the benevolent and sovereign whose blessing and glory flourished my thoughts and thrived my ambitions. Trembling lips and wet eyes praise for HOLY PROPHET (P.B.U.H) for enlightening our conscience with the essence of faith in ALLAH, converging all His kindness and mercy upon Him. I would like to take this opportunity to express my gratitude to my supervisor DR. Muamer N. Mohammed and my co-supervisor Dr. Kohbalan A/L Moorthy for their constant guidance and encouragement

I cannot finish my acknowledgements without thanking my family for their affectionate understanding and love throughout my PhD. study; my father, my mother, my brothers, and my sisters. No words could express my deepest gratitude for my family, Their support and efforts have relieved my worries and made me able to concentrate on my research, They have always encouraged me when I encountered difficulties. I would like to share this moment of happiness with them. Special thanks with all my heart to my wife for being the greatest support during my PhD life, which makes my time towards the PhD program as smooth as possible.

Last but not least, I would like to thank everybody who was important to the successful realization of the thesis, as well as express my apology that I could not mention personally one by one. The completion of my dissertation would not be possible without the help and encouragement of many people.

ABSTRAK

Ledakan pertambahan bilangan imej biometrik yang disimpan di dalam kebanyakan pangkalan data telah menjadikan pengelasan imej sesuatu yang mandatori. Proses sebegini boleh mempengaruhi kelajuan capaian data, selain menyokong proses pengambilan semula. Para penyelidik kini menumpukan perhatian terhadap usaha mengenal pasti ciri-ciri imej yang sesuai digunakan untuk pengelompokan dan pengindeksan dengan proses carian yang efisien. Kaedah tersedia tidak mampu mengekstrak bilangan ciri terpenting imej iris yang mencukupi untuk proses pengelompokan dan pengindeksan. Namun begitu, salah satu kelemahan pengelompokan ialah proses mengekstrak ciri-ciri terpenting. Suatu gabungan tiga kaedah transformasi iaitu Transformasi Kosinus Diskret (Discrete Cosine Transformation, DCT), Transformasi Gelombang Diskret (Discrete Wavelet Transform, DWT) dan Penguraian Nilai Tunggal (Singular Value Decomposition, SVD) untuk menganalisis imej iris dan mengekstrak ciri-ciri setempatnya belum pernah digunakan untuk pengelompokan dan pengindeksan imej. Masalah lain berkaitan pengelompokan ialah ketika memilih sentroid awal secara rawak untuk setiap kelompok. Kelemahan ini diatasi menggunakan Algoritma Kunang-kunang (Firefly Algorithm, FA) kerana ia mampu melaksanakan carian global dan mempunyai kadar penumpuan pantas untuk mengoptimumkan pusat pengelompokan awal algoritma K-purata (K-means algorithm), menggunakan sejenis jarak Euclid terwajar untuk mengurangkan kecacatan akibat data hingar dan lain-lain ketidakpastian. Tesis ini membentangkan suatu kaedah baru untuk mengekstrak ciri paling sesuai daripada imej biometrik iris untuk mengindeks pangkalan data dalam tempoh dan kawasan carian yang minimum. Kaedah dipertingkatkan ini menggabungkan tiga kaedah transformasi untuk menganalisis imej iris dan mengekstrak ciri-ciri setempatnya. Kaedah ini menggunakan algoritma pengelompokan K-purata terwajar berasaskan FA diperbaik untuk mengoptimumkan pusat pengelompokan awal algoritma K-purata, yang dikenali sebagai Algoritma Pengelompokan K-purata Terwajar-Algoritma Kunang-kunang Diperbaik (Weighted K-means clustering-Improved Firefly Algorithm, WKIFA). Bertujuan carian dan ambilan semula, suatu teknik selari cekap dibentangkan dengan membahagi kumpulan ciri-ciri kepada dua pohon-b berdasarkan kunci indeks. Carian dalam suatu kumpulan boleh dilakukan menggunakan algoritma carian separuh untuk meningkatkan masa tindak balas untuk pengambilan semula data. Sistem ini diuji menggunakan pangkalan data umum. Dapatan kajian menunjukkan bahawa sistem pengindeksan ini mempunyai kadar penembusan yang agak rendah, iaitu pada 0.98%, 0.13% dan 0.12% dan kadar tersasar tong yang rendah pada 0.3037%, 0.4226% dan 0.2019% berbanding pangkalan data iris tersedia masing-masing, milik Akademi Sains - Institut Automasi China (CASIA), Universiti Bath (BATH) dan Pangkalan Data Institut Teknologi Kanpur, India (IITK). Dapatan kajian untuk WKIFA diperbaik menunjukkan bahawa kaedah ini lebih berkesan untuk peringkat pengelompokan sistem. Malah, ia melebihi prestasi K-purata tradisional dengan mengurangkan kadar penembusan kepada 0.131%, 0.088% dan 0.108% dan meningkatkan ketepatan dengan mengurangkan kadar tersasar tong kepada 0.2604%, 0.309% dan 0.1548%, masing-masing untuk pangkalan data yang dinyatakan terdahulu. Analisis kerumitan masa pengambilan semula pula menunjukkan bahawa kerumitan pengiraan dikurangkan kepada O (log N), iaitu lebih baik berbanding kaedah sedia ada.

ABSTRACT

The explosive increase in the number of biometric images saved in most databases has made image indexing mandatory. These processes could influence the speed of data access as well as support their retrieval. Hence, researchers are focusing on how to determine suitable image features to be used for clustering and index, with an efficient searching process. The existing methods are unable to extract sufficient number of the most important features of iris image for clustering and indexing processes. However, one of the weaknesses of clustering is the process of extracting the most important features. A combination of three transformation methods, namely, Discrete Cosine Transformation (DCT), Discrete Wavelet Transform (DWT), and Singular Value Decomposition (SVD) for analyzing the iris image and for extracting its local features have yet to be utilized for image clustering and indexing. Another problem related to clustering is when choosing the initial centroids for each cluster randomly. To overcome this disadvantage, the Fireflies Algorithm (FA) was used because it has the ability to perform global searches and has quick convergence rate to optimize the initial clustering centers of the K-means algorithm, using a kind of weighted Euclidean distance to reduce the defects made by noise data and other uncertainties. This thesis presents a new method to extract the most relevant features of iris biometric images for indexing the database within minimum time and search area. The enhanced method combines three transformation methods for analyzing the iris image and extracting its local features. It uses a weighted K-means clustering algorithm based on the improved FA to optimize the initial clustering centers of K-means algorithm, known as Weighted K-means clustering-Improved Firefly Algorithm (WKIFA). For searches and retrieval, an efficient parallel technique has been presented by dividing the group of features into two b-trees based on index keys. Searches within a group can be done using a half-searching algorithm to improve the response time for data retrieval. The system has been tested on publicly available databases. The experimental results showed that the indexing system has a considerably low penetration rate of 0.98%, 0.13%, and 0.12%, and lower bin miss rate of 0.3037%, 0.4226%, and 0.2019% compared to the existing iris databases of the Chinese Academy of Science - Institute of Automation (CASIA), University of Bath (BATH), and Database of Indian Institute of Technology Kanpur (IITK), respectively. Results of the improved WKIFA showed that it was more effective for the clustering stage of the system. It even outperformed the traditional K-mean, by reducing the penetration rates to 0.131%, 0.088%, and 0.108%, and improving the accuracy by reducing the bin miss rate to 0.2604%, 0.309%, and 0.1548% of the aforementioned databases, respectively. Analysis of time complexity of retrieval showed that the computational complexity was reduced to O (log n), which was better than the existing methods.

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

σ	$M \times N$ Diagonal Matrix
$D\ (x)$	Distance with Nearest Cluster
$O(1)$	Logarithmic Time
$T\ (n)$	Frequency of Time
U	$M \times M$ Orthogonal Matrix
V	$N \times N$ Orthogonal Matrix
σ_1	First Singular Value
σ_2	Second Singular Value
$\sum\dot{D}$	lower-frequency coefficient
μ	average of all intensity value
θx	probability of points
D(x)	distance to center
f(xi)	objective function in firefly
r	fluorescence brightness
γ	light intensity
β	attraction
E_i	value of energy
$F_{i,j}$	dimensional of features
X_{id}	first component
ω	overall distribution of the sample data
Γ_i	number of data in the cluster
f(n)	the number of comparisons required
C^i	the candidate set
T_{max}	Maximum number of iterations
$O(1)$	constant in in complexity theory
$O(\log(n))$	logarithmic in in complexity theory
$O((\log(n))^c)$	polylogarithmic in in complexity theory
$O(n)$	linear in in complexity theory
$O(n^2)$	quadratic in in complexity theory
$O(n^c)$	polynomial in in complexity theory
$O(c^n)$	exponential in in complexity theory

LIST OF ABBREVIATIONS

AES	Advanced Encryption Standard
AHE	Adaptive Histogram Equalization
AR	Accuracy Rate
ATM	Automated Teller Machines
BATH	University of Bath Iris Image Database
BM	Rate of Bin Miss
BWT	Borrows-Wheeler Transform
CASIA	Chinese Academy of Sciences Iris Image Database
CLAHE	Contrast Limited Adaptive Histogram Equalization
CM	Confusion Matrix
CMC	Curve of Cumulative Match Characteristic
DB	Database
B-tree	Binary Tree
DCT	Discrete Cosine Transform
DFT	Discrete Fourier Transform
DWT	Discrete Wavelets Transform
EER	Equal Error Rate
FAR	False Acceptance Rate
FBI	Federal Bureau of Investigation
FMR	False Match Rate
FNMR	False Not Match Rate
FRR	False Rejection Rate
FTC	Failure to Capture
FTE	Failure to Enroll
GAR	Genuine Acceptance Rate
HFP	High Frequency Power
IAFIS	Fingerprint Identification System
IBA	International Biometric Association
IBIA	International Biometric Industry Association
ICE	Iris Challenge Evaluation Iris Image Database
ID	Identification

IDCT	Inverse Discrete Cosine Transform
IITK	Indian Institute of Technology Kanpur
IS	Identification Services
LBP	Local Binary Pattern
MMU1	Multimedia University Iris Image Database
NBSP	National Biometric Security Project
NGI	Next Generation Fbi Iafis
NIR	Near Infrared
NIST	National Institute of Standards And Technology
PCA	Principle Component Analysis
PDA	Personal Digital Assistant
PR	Penetration Coefficient
PRNG	Peak Signal-To-Noise Ratio
ROC	Curve of Receiver Operating Characteristic
ROI	Region of Interest
SIFT	Scale Invariant Feature Transform
SPLDH	Signed Pixel Level Difference Histogram
SURF	Speed Up Robust Features
SV	Singular Values
UBIRIS	University of Beira Interior Iris Image Database
UIDAI	Unique Identification Authority of India
UPOL	University of Palackeho And Olomouc
WVU	West Virginia University Iris Image Database
WKIFA	weighted K-means based on improved Firefly
FR	firefly algorithm

CHAPTER 1

INTRODUCTION

1.1 Background of Study

In an increasingly digital world, reliable personal authentication is an important human computer interface activity. As such, biometrics such as iris verification are gaining industrial, government and citizen acceptance, because of its numerous benefits over other biometric traits (Emad & Norrozila, 2015) as presented in Figure 1.1. Notably, iris biometrics has facilitated the possibility for generating various large scale real time databases worldwide. For instance, the United Arab Emirates (UAE) has launched a national border crossing security initiative. Presently 27 land, air and sea ports of entry are equipped with this system (Mehrotra & Majhi, 2013).

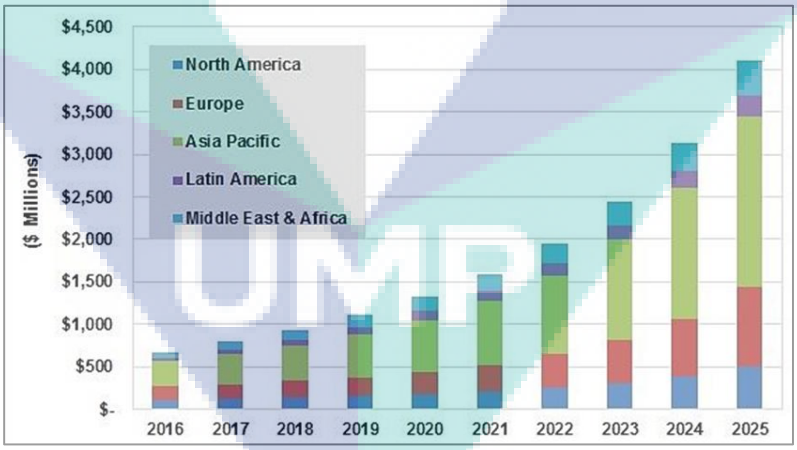


Figure 1.1 Annual iris recognition revenue by region markets between 2016 and 2025

Source: Tractica.com (2017).

In India, a large scale project Aadhaar has been put in place to issue unique identification number to each individual using fingerprint and iris (Kavati *et al.*, 2017; Unique Identification Authority of India, 2017). In the UK, iris recognition immigration system (IRIS) is being used to enter through automated barriers at certain airports. However, the biometric system faces the scalability issue as the number of people to be enrolled into the system runs into billions. As such, this issue has become the bottleneck due to low response time, as against the desired high search and retrieval efficiency in addition to accuracy (Dey & Samanta, 2012). Unfortunately, despite the rapid proliferation of large-scale databases, the research community has thus far focused only on accuracy within small databases while neglecting the scalability and speed issues which are more important to large-scale applications. Hence, the number of false acceptances grows geometrically with increase in the size of the database (Mehrotra & Majhi, 2013; Pyykkö, 2018). It is well known that the time required claiming identification is directly proportional to the database size. Thus, there is a stringent requirement to partition the gallery set so as to compare the probe iris with only those elements of gallery set that possesses similar characteristics. However, this can be achieved with the help of some classification and indexing approaches. The classification scheme partitions the database into some supervised classes. As such, the class of probe iris can be firstly estimated and compared only against subset of identities of gallery set that belongs to the probe class.

1.2 Problem Statement

With the increasing size of biometric databases, reliability and scalability issues have become a challenge to accuracy, low response time, high search and high retrieval efficiency (Kavati *et al.,* 2016). Hence effort is being concerted towards reduction of the search space, in a bid to improve the performance of biometric systems. However, this requires an efficient indexing and clustering scheme. Clustering is saddled with shouldering the responsibility of determining the success and efficient partitioning and clustering of the system. Notwithstanding, there are two aspects of clustering related problem - extraction and clustering (Raykov *et al.*, 2016). The first aspect of the problem is the ambiguity of extracting important and sufficient numbers of features from biometric to distinguish among different biometric and the relations between them.

The efficiency of clustering processes depends on the strength of the image clustering algorithm. The commonest clustering methods are hierarchical and partitioning clustering. With the partitioning clustering method, all the clusters are simultaneously found without the need of forming a hierarchical structure. One of the main partitioning clustering approaches is the centre-based clustering approach. Literature review shows that the k-means algorithm is one of the commonly used partitioning clustering algorithms (Jain 2010; Meila & Heckerman 2013; Raykov *et al.,* 2016). However, despite the fact that the k-means can be effectively used in many cases, it is sensitive to noise and outlier points. Unfortunately, even a few such points can have a significant influence on the means of their respective clusters (Celebi *et al.,* 2013). Additionally, the performance of k-means can be affected by the choice of initial values that may converge to the local optima of the criterion function instead of converging at the global optima (Jain 2010; Meila & Heckerman 2013; Raykov & Little, 2016). It may further generate empty clusters, high probability of local optima stagnation and slow convergence when improperly implemented (Celebi, 2011; Jain, 2010; Meila & Heckerman, 2013). Notwithstanding, the k-means has been shown as the best clustering algorithm despite its failure in overcoming performance challenges (Bouras *et al.,* 2010; Jain 2010; Meila and Heckerman, 2013).

The identification in the biometric systems is usually done sequentially by comparing the query biometric against every enrolled individual's template in the database. This process is computationally expensive and increases the response time of the system (Kavati *et al.,* 2015). Besides, reducing the search space is not enough solution for high dimensional data as most of the partitioning and indexing methods have neglected the required computational efforts of searching data (Rathgeb *et al.,* 2015; Dey, & Samanta, 2012). However researchers still did not use approach of searching inside each created group of clusters to retrieve the information.

1.3 Research Objectives

The objectives of this study are stated below:

i. To improve the efficiency and accuracy of indexing technique by extracting sufficient numbers of the most important features of iris image for clustering and indexing process, so as to use it for clustering and indexing the biometric database. by combined three popular transformation methods: The DCT and DWT to take their advantages Individually and collectively, then SVD transform is used to reduce the amount of data as well as brings out the useful part of the data to extracted them as the most relevant features of image to be used for clustering and indexing using the scalable K-means++ Algorithm.

ii. To improve the accuracy of clustering process in order to reduce the defects made by noise data and other uncertainties, as well as to optimize the initial clustering centres of traditional K-means algorithm, by presenting an enhanced clustering algorithm (WKIFA), which is based on added a kind of weighted to Euclidean distance and combine Firefly Algorithm after improved. Fireflies Algorithm (FA) which has power ability of global search and quick convergence rate.

iii. To improve the data retrieval time of the enhanced iris biometric system by developing an efficient searching approach inside each created group based on two B-tree structure, parallel searching and half-searching algorithm. The feature groups are separated and distributed into two databases, both of which are organised as B-tree based on the global index key of the images. The search for the query image is carried out by the global key, which is used to traverse a node of the tree to reach the specific group and half search method used to search for and retrieve the candidates inside each bin based on their similarity.

iv. To evaluate the performance of the improved classified and indexing system in terms of accuracy as well as the retrieving time, and also to verify the validity and feasibility of the new clustering algorithm.

1.4 Scopes of Research

The scope of the current study is to solve the issues emerging from the extraction of features based on hybrid three transformation methods which are Discrete Cosine Transformation (DCT), Discrete Wavelet Transform (DWT), and Singular Value Decomposition (SVD). In addition, the study extends to solve the weakness of the k-means algorithm using an improved clustering algorithm which depends on a weighted K-means clustering algorithm based on the improved Firefly Algorithm (WKIFA). The study strives to present an efficient searching approach that will reduce the image retrieving and search time. The developed approach in this study is based on a B-tree structure using parallel searching and half-searching algorithms. Many benchmark datasets were used in this research from two categories: Cooperative databases and Non-cooperative Databases, the database for eye image was downloaded from: the Chinese Academy of Science - Institute of Automation (CASIA), University of Bath (BATH) and Indian Institute of Technology Kanpur (IITK). It should be noted that the iris image acquisition that was performed by capturing eye images was not be covered in this project. Iris is the only biometric source used for feature extraction in this work. Fusion and indexing are not covered in this study, where the performance of any feature extraction process can be enhanced by deploying various sources of evidence (multimodal biometrics).

1.5 Significance of Study

The primary motivation for enhancing indexing and retrieving of biometric systems is to improve the response time in iris biometric systems from large databases. The comparisons in a large database does not only increase the data retrieval time but also decreases the error rates. As a solution, it is suggested that each biometric template in an indexing database should be assigned an index value to reduce the systems' search space (Kavati, Prasad, & Bhagvati, 2017). This should be done using the most relevant features as the extracted features are important for successful clustering and indexing processes; it should also be highly relevant, as well as correctly represent the biometric image (Singh Patwal & Srivastava, 2016). On the other hand, clustering is an important step in the enhanced indexing system. Clustering refers to the partitioning of data objects into groups (clusters) with the aim of enhancing the clustering algorithm for better partitioning accuracy. Likewise, an efficient searching approach is another motivation because it can be a slow process since each input biometric data must be sequentially matched against all the biometric data in the system. The rate of false acceptance error often grows with the size of the database, as a result, the systems' response time, search and retrieval efficiency, as well as identification accuracy, might be affected (Parmar & Degadwala, 2015).

1.6 Thesis Layout

The thesis is structured into four chapters. The chapter 1 presented a detailed background overview of this study which includes the problem statement, research objectives, scopes, and significance of study. Moreover, the chapter 2 contains definitions, explanation about biometric and system design, the properties of iris biometric and database, transformation and clustering techniques as well as measures of performances. Also, the chapter 3 presents the enhanced techniques and the developments to the indexing and classification system, as well as the improved searching and retrieving approach for the system. In chapter 4 the experimental results and discussions of the improved schemes are explained. On the other hand, Chapter 5 presents the conclusion and recommendations for future work. Please refer to Figure 1.2.

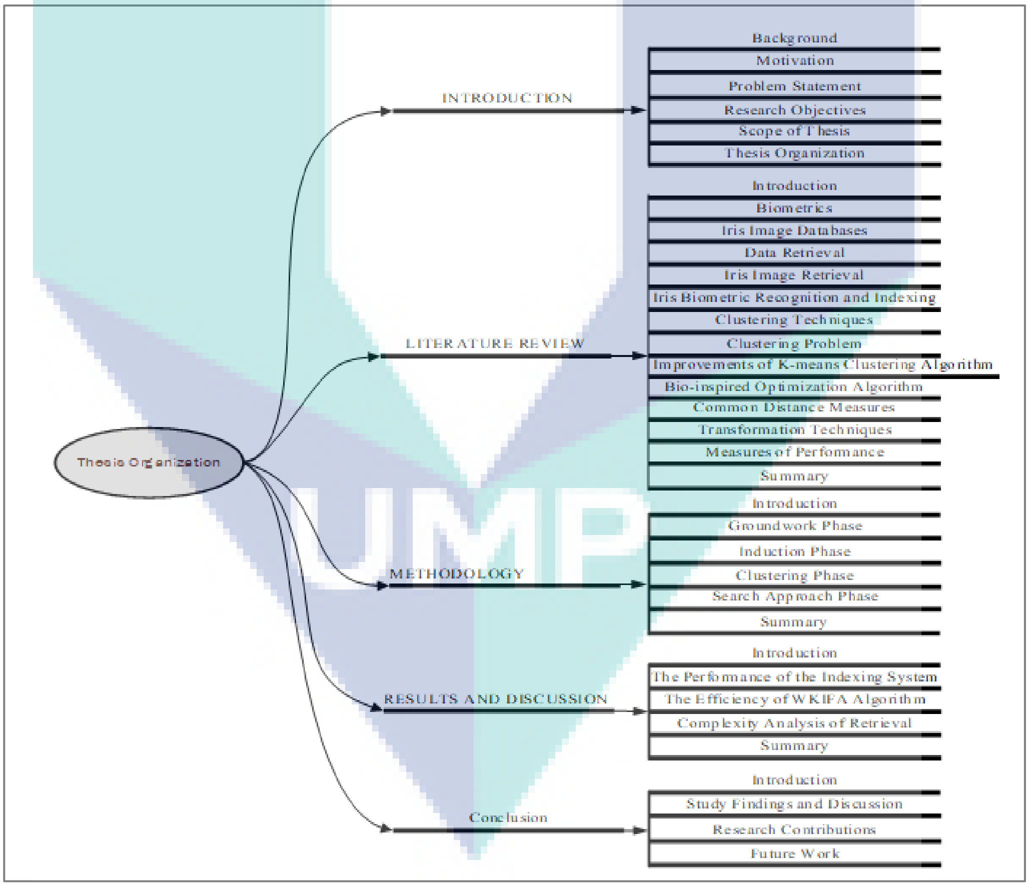


Figure 1.2 Thesis organization

CHAPTER 2

LITERATURE REVIEW

2.1 Overview

Biometric identification is the process of associating an identity to the input biometric data by comparing with other existing identities that are available in the database to find a matching template for identification (Kavati *et al.*, 2015). In many systems, this comparison is undertaken in an exhaustive manner, i.e., the input data of an individual is compared against all enrolled data in order to determine the identity of the individual. However, due to the large number of entries in the database, one to one matching of the query iris with each iris in the database would be computationally infeasible, turns into a bottleneck for low response time, affecting the efficiency of high search and retrieval, as well as accuracy. Therefore, a filtering process is usually invoked in order to reduce the number of candidate hypotheses for matching operation. Filtering can be achieved by two difierent approaches: classification and indexing. Furthermore, an effective searching method is needed if the database has a huge amount of biometric templates/data, this thesis focuses on the iris-based biometric systems due to their numerous benefits over other biometric traits (Tractica.com, 2017; Khalaf *et al.*, 2016). Generally, the iris refers to the protected organ that is visible on the external with permanent and distinctive epigenetic pattern of a person, which is perfect to be utilized for recognizing a person. With that, some methods that process the images captured can be used to extract the distinctive pattern of iris from the digital eye image to be encoded as biometric data/template for storage in database. This particular biometric template consists of an objective representation of mathematical feature concerning the special data extracted from the iris.

2.2 Principles of Biometrics

Biometrics is a science of automatically identifying an individual based on their unique physiological or behavioural characteristics. Sir William Herschel demonstrated an initial use of handprint biometrics to verify worker‘s identity in 1858 (NTSC). In mid-19th century, Alphonse Bertillion, a law enforcement officer was famous for using physical body measurements to identify criminals (Chaudhari *et al.*, 2013). In late 19th century, use of fingerprints for person identification was proposed. Law enforcement departments immediately embraced the idea of using fingerprints for person identification based on the traits distinctiveness. However, initially, person identification based on fingerprints was manually verified.

In early 20th century automated fingerprint biometric systems for person recognition were introduced (Alsmirat *et al.*, 2018). Since the introduction of automated biometric systems, use of biometrics has seen an exponential growth. Automated biometric system or biometric system is a device that collects a biometric sample from an individual, extracts features, compares the features against the available templates in the database, and returns a recognition result. Some of the most common biometric systems use fingerprints, face, iris and, voice traits (Taha & Norrozila, 2015). The design of all the biometric systems comprises of a two-step process of enrolments and recognition (Figure 2.1).

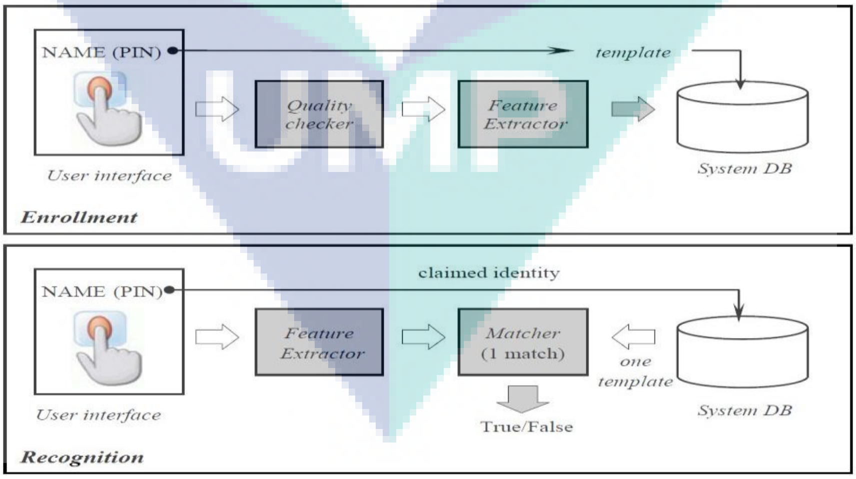


Figure 2.1 Biometric models to authenticate

Source: Fierrez *et al.*, (2018)

The enrolment process involves scanning the biometric traits of an individual on the biometric system for further processing to extract the unique representations of the individual (template). The extracted template of the individual can then be stored in a local or central database. The recognition process is initiated when an enrolled user wants to confirm his identity to gain access. During the recognition process, the biometric trait of the individual is scanned and processed in a similar way as in the enrolments phase. The processed information is then matched to an existing template in the database to confirm the identity of the individual.

Biometric systems can be used in either verification or identification mode depending on the application. When used in the verification mode, the obtained biometric trait from an individual is matched with his/her own templates which have initially being stored in the systems' database. This verification mode is often called one vs. one comparison as the user identification is based on his/her claimed identity. When used in the identification mode, the biometric trait obtained from a person is matched against several templates to confirm an identity. This method is often called one vs. many comparisons as the user identification is based on comparison with many templates. Biometric traits can be classified into physical and behavioural traits All biometrics systems are designed in a two-step process: Enrolments process and Recognition process. In enrolments process, the biometric system scans the biometric trait of an individual and further processes the trait to extract compact and meaningful representation. This representation of an individual‘s biometric trait is known as a template. The biometric system then stores the template in a local/central database. Recognition process is activated when the user wants to claim his/her identity to gain access. In the recognition process, the biometric system scans the biometric trait and processes the trait in a similar fashion as in enrolments phase. The processed information is then compared to existing template in the database to establish the identity. Depending on the application, biometric systems can be realized in two modes: verification mode and identification mode. In verification mode, the system matches the biometric sample obtained from the individual with his/her own biometric templates stored in the system database.

The verification mode is generally known as one vs. one comparison, where a user is identified based on a claimed identity. In identification mode, the obtained biometric sample from an individual is matched against various templates (many users) to establish an identity. This is commonly known as one vs. many comparisons, where a user is identified without Biometric traits can be broadly divided into physical and behavioural traits (Taha & Norrozila, 2015). Physical traits are iris, fingerprint, face, eye vasculature; retinal vasculature, DNA, eye shape etc. and examples of behavioural traits are speech, signature, handwriting, gesture etc. Physical traits are something that can be measured over time. Behavioural traits are generally learned over time and are acquired after an effort from user. Voice biometric is one trait that is both physical and behavioural. Voice biometric is measured using vibrations of vocal cords and vocal tract shape, but also depends on user‘s behavioural state such as state of mind.

2.2.1 Design of Biometric System

Basically, the concept of identification biometric system is comprised of four essential steps, the fundamental blocks of building for generic biometric application is shown in Figure 2.2. The first time an individual uses a biometric system is called an enrollment. During the enrollment, biometric information from an individual is stored. In subsequent uses, biometric information is detected and compared with the information stored at the time of enrollment.

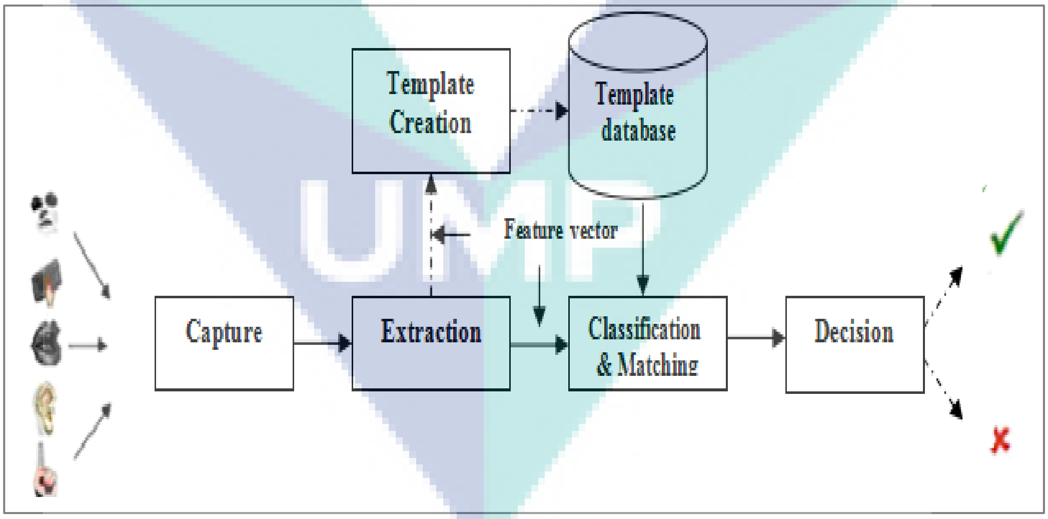


Figure 2.2 Fundamental blocks of building for generic biometric application

Source: Mazumdar & Nirmala, (2018)

Note that it is crucial that storage and retrieval of such systems themselves be secure if the biometric system is to be robust. The first block (sensor) is the interface between the real world and the system; it has to acquire all the necessary data. Most of the times it is an image acquisition system, but it can change according to the characteristics desired. The second block performs all the necessary pre-processing: it has to remove artifacts from the sensor, to enhance the input (e.g. removing background noise), to use some kind of normalization, etc. In the third block necessary features are extracted.

This step is an important step as the correct features need to be extracted in the optimal way. A vector of numbers or an image with particular properties is used to create a template. A template is a synthesis of the relevant characteristics extracted from the source. Elements of the biometric measurement that are not used in the comparison algorithm are discarded in the template to reduce the file size and to protect the identity of the enrollee (Mazumdar & Nirmala, 2018). If enrolments is being performed, the template is simply stored somewhere (on a card or within a database or both). If a matching phase is being performed, the obtained template is passed to a matcher that compares it with other existing templates, estimating the distance between them using any algorithm (e.g. Euclidean distance). The matching program will analyse the template with the input. This will then be output for any specified use or purpose (e.g. entrance in a restricted area) (Mazumdar & Nirmala, 2018).

2.2.1.1 Data Requirement

The incorporation of viable interface for users in the application of sensor of biometric/reader is significant, especially in measuring or gathering organic data from biometrics using any device, such as digital camera, sensor, scanner, etc. Besides, being in between the actual world and the system, data are captured from any part to identify a person. Moreover, data requirement is indeed significant because the quality of raw biometrics heavily relies on the characteristics of the device that capture the images (Vielhauer, 2011).

2.2.1.2 Feature Extraction

Generally, the required raw biometric data need to undergo some pre-processing operations before the features can be extracted. In fact, this particular step is rather significant especially to decide the trait to be extracted and the method that has to be employed. In short, feature extraction reflects the procedure of retrieving a concise image of expressive digital quality for a predetermined trait biometric (known as a template). This template comprises of the information obtained from biometric samples and are used to generate fresh biometric traits known as feature sets. As such, these traits must be distinct in nature for the individual, besides being invariant to alteration in various users for the similar trait biometric of the similar individual (minute intra-user variance). Besides, the set of attributes retrieved at the enrolment phase is kept in the system database as a template. This carries out the necessary pre-processing operations - the removal of artefacts from the data, the enhancement of the input after some processes (removal of noise), and the use of some normalization effects (Easwaramoorthy *et al.*, 2016; Vielhauer, 2011).

2.2.2 Properties of Biometrics

The application of biometrics enables the establishment of the identity of an individual depending on the person itself rather than possessions like ID cards or keywords to be remembered (Blasco *et al.*, 2016). Therefore, a question arises: ‘What are the qualified measurement of physiological traits needed to establish biometrics?’ Hence, numerous traits can be implemented for biometric features only if the following demands are met: 1) Universality, 2) Uniqueness, 3) Distinctiveness,4) Permanence,5) Collectability, 6) Acceptability, 7) Performance, 8) Circumvention.

2.3 Iris Biometrics

In 1987, the automated biometrics system was first initiated by Flom and Safir (Ansari & Aquib, 2017). On the other hand, the iris biometric application was initiated by Daugman at the Cambridge University (Ansari & Aquib, 2017) where shots of iris were taken by utilizing a source of near-infrared light due to its ability in controlling radiance, besides being safe for users. The next process determines if the captured iris image is a deformable datum trained using certain attributes; the eye shape was highlighted for the process of authentication (Ansari & Aquib, 2017). Other than that, Daugman presumed that both pupil boundary and iris are in circular shape; hence, dictated by three essential features: radius, r; center of the circle, x_0; and y_0 (Saad & George, 2014). Meanwhile, reliability is heavily dependent on the capability of gaining distinct traits which can be obtained over time (Saad & George, 2014). Furthermore, although each biometrics has some advantages and disadvantages, their application relies on certain settings. For instance, the fingerprint trait has remained unique since way long back in time, while distinct face features can differ considerably with variances in time and place. Moreover, fewer limitations must be imposed upon users who provide data for biometric applications. On top of that, fingerprint acquisition is an invasive method for it only needs a user to touch the sensor. A block diagram of the iris biometric system is shown in Figure 2.3.

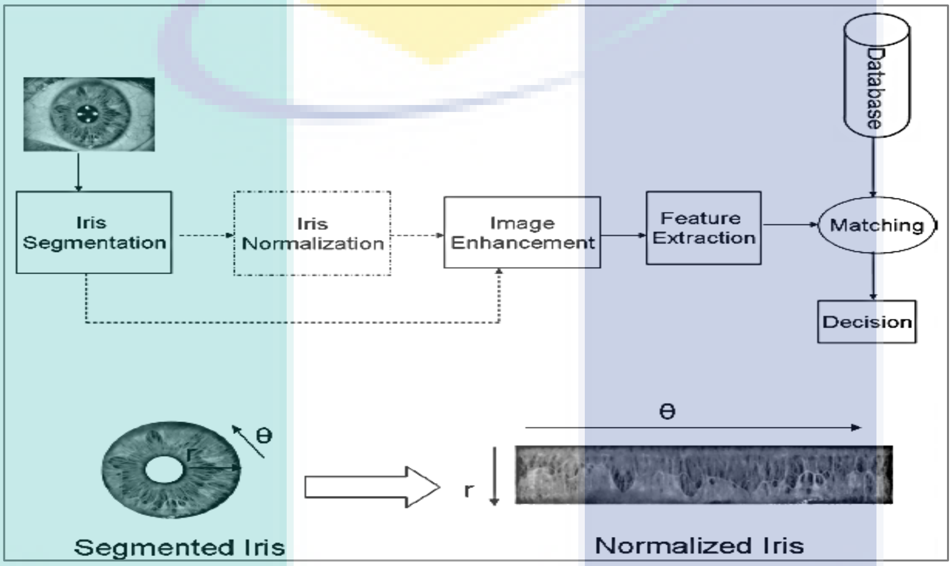


Figure 2.3 Iris biometric system

Source: Ganorkar & Rahman, (2013)

Among the many biometric features, the iris is deemed as important because it is used to verify a person by using distinct texture patterns (Hamd, Ahmed, 2018). Besides, by weighing in consistency and insensitivity, iris has been found to be the most effective method. Besides, as for consistency, patterns of spatial must be distinct between users, while for insensitivity, the iris should not change being an inner organ that is protected with a texture that is random. Hence, iris can be adapted to be a secret code, which cannot be forgotten but embedded within the user. Moreover, such a system offers simultaneous and high confidence identification for users via a mathematical approach for visible structures of eye iris within a certain space. The aspect of randomness for the iris patterns reflects very high dimensionality, besides being one of the most reliable biometric features at hand. Hence, authentication decisions can be made with high levels of confidence, sufficient to upkeep quick, reliable, and thorough quests combed datasets from the national level. In addition, iris has been deemed as the most significant part in an eye image as shown in Figure 2.4.

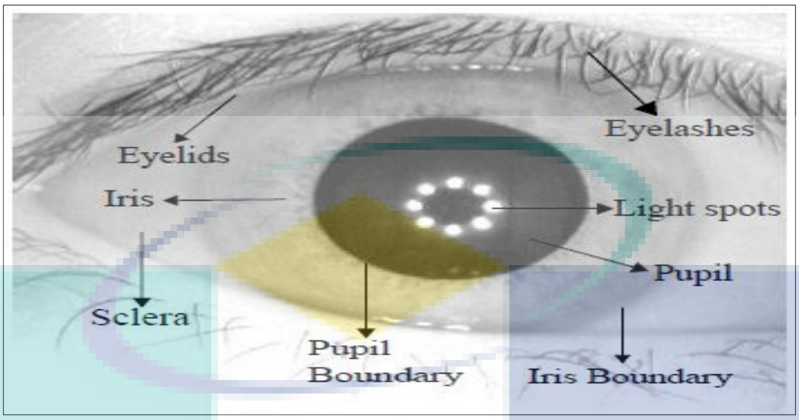


Figure 2.4 Image #: S1001R01 from CASIA database shows eye anatomy

Source: Biometrics Ideal Test, (2017)

The iris, which is a circular ring in shape, consists of numerous overlapping minute features, such as freckles, coronas, stripes, furrows, and crypts, to name a few. These tiny iris patterns are distinct to each person and non-invasive. As one probe into the iris, a central dark circle known as the pupil is found. In fact, the iris possesses muscles that constrict the pupil in brighter light and dilate when dimmer. Besides, the amount of light entering the eye is controlled by this pupillary motion. On top of that, the pupil circumference and the iris are also defined as pupil and iris boundary, respectively. Meanwhile, sclera refers to the white area, which consists of tissues that are tough and leather-like around iris. Besides, the upper and lower eyelids cover the eyeball. The upper eyelid refers to the membrane that is stretchable and covers the eye, generating a movement that ranges from close to open widely. Lower eyelid, nonetheless, possesses restricted movement because of the position of the eyeball (Bernard, 2013). Other than that, hair protruding from the eyelid edge, which is known as eyelashes, prevents dust from entering into the eye. Hence, the image processing methods are applied to detect a special pattern in iris identified in the shot images of eyes for storage.

Iris, which refers to the inner eye organ, is found in front of the lens, but behind the cornea and aqueous humor, Besides, Iris can be said to be the one and only visible inner organ which is external. Iris images meant for personal recognition are captured as far as approximately 100 cm (Saad & George, 2014). The visible traits of the iris include the ciliary processes, the contraction furrows, the trabecular meshwork of connective tissue, the collagenous stromal tissues, the crypts, a corona and pupillary frill, rings, freckles, and coloration. The predominant texture that can be seen with visible light is generated by the striated anterior layer that covered the trabecular meshwork, as presented in Figure 2.5.

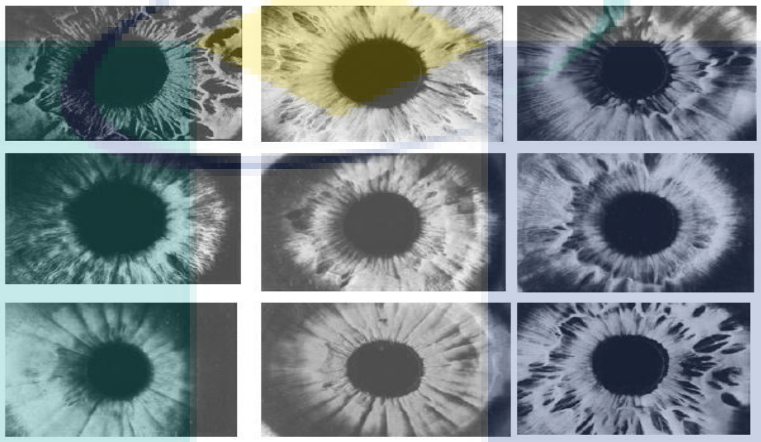


Figure 2.5 Examples of visibly different iris patterns

Source: Fouad, (2012)

In fact, all these sources of radial and angular variations, when considered together, form a distinctive "iris print", which can be imaged from some distance. Other significant features of the iris that determine its adequacy to be used in authentication application are given in the following: (i) its isolation and protection from the external environment; (ii) surgically altering it is impossible without imposing risk to vision; and (iii) its physiological response to light that offers natural tests against artifice (Diego *et al.*, 2015). Moreover, iris and fingerprints share a common feature, which is the random morphogenesis of its minutiae. Besides, as the genetic aspect is absent from iris beyond its anatomical form, colour, physiology, and general appearance; iris has a texture that is stochastic or chaotic.

In addition, since the detailed morphogenesis heavily relies on the initial phases of the embryonic mesoderm from which it develops, the phenotypic expression even of two irises with similar genetic genotype (as in identical twins, or the pair possessed by one individual) has uncorrelated minutiae. Hence, iris is indeed as unique as fingerprint, disregarding if they share similar genotypes. On top of that, iris has several benefits over other biometrics for automatic authentication, such as: (i) the ease of capturing its image at some distance from a subject without physical contact, in an un-instructive manner, and perhaps discreetly; (ii) its fundamental polar geometry that conveys a natural coordinate system and coordinates; as well as (iii) the highly random patterns, creating inter-subject variability that spans from 266 degrees-of-freedom (Daugman, 2006). Furthermore, the human iris starts to develop at the third month of gestation. The structures, nonetheless, generate a distinctive pattern that completes by the eighth month of gestation. However, pigmentation continues into the first year after birth, when the chromophore cells shift the colour of the iris, although clinical evidence asserts that the trabecular pattern is stable throughout one's lifespan (Smartsensores, 2014).

Meanwhile, the layers of the iris are comprised of ectodermal and mesodermal embryological origins, which consist of (from back to front): pupillary dilator and sphincter muscles; a darkly pigmented epithelium; heavily vascularized stromal (connective tissue of interlacing ligaments containing melanocytes); and an anterior layer of chromataphores and melanocytes with genetically determined density of melanin pigment granules. Hence, the combined effect is a visible pattern that projects numerous unique traits, such as crypts, arching ligaments, ridges, furrows, and a zigzag collarets. As for the iris colour, it is mainly determined by the density of the stromal and its melanin content, while blue irises surface due to absence of pigment: long wavelength light penetrates and it is absorbed by the pigment epithelium, whereas shorter wavelengths are reflected and scattered by the stromal. Even though the heritability and ethnographic diversity of iris colour have been looked into by many, only a handful have investigated the intricate achromatic pattern and the textural variability of the iris among people (Iris Challenge Evaluation (ICE), 2017). Other than that, since it is an internal organ of the eye, the iris is immune (unlike fingerprints) to environmental impacts, except for its pupillary response to light.

Furthermore, the elastic distortions that take place with pupillary dilation and constriction are reversed mathematically by using algorithms in order to localize both the inner and the outer boundaries of the iris. The pupillary motion, even without any changes in illumination (termed "hippus"), as well as the related elastic distortion in the iris texture, offers test against photographs, glass eyes, or other simulacra for a living iris. Nevertheless, other tests incorporate changing infrared LED light sources that could cause some alteration in the specular reflections from the cornea; determining the attributes of the contact lens that may have a printed fake iris pattern on the spherical surface of the cornea, rather than in an internal plane in the eye; as well as determining the attributes of living tissue under various wavelengths of both visible and infrared illumination.

2.4 Iris Image Databases

The biometric process comprises of an automated method that authenticates individuals based on their biological traits. Since the past thirty years, the field of biometric recognition has gained interest among many researchers, and several proposals have been introduced. While these proposals are waiting to be proved, a large number of tests over a large number of subjects have to be performed. Therefore, it is not realistic for researchers to collect their own dataset due to the many challenges one might face in doing so. Besides, for a more robust comparison, different proposals should be implemented over the same dataset. Therefore, a benchmark database has been deemed as necessary and cannot be sacrificed for the sake of recognition development (Alrifaee *et al.*, 2017).

2.4.1 CASIA Iris Database

This is the freely available iris database for researcher purposes. It was developed by the Chinese Academy of Science (CAS) - Institute of Automation (Center for Biometrics and Security Research, 2017; Alrifaee *et al.*, 2017). The initial versions of the CASIA iris databases had utilized optimum capture surroundings with stop and stare at near proximities. The researchers also used NIR (near infrared) light sources which shared similar conditions with those implemented in Daugman’s iris identification system. From the time of developing the CASIA database, the CAS has formed 4 variants of the CASIA as given below in the proceeding sections.

2.4.1.1 CASIA-Iris V1

The CASIA-Iris version 1 comprises of 756 images of iris captured from 108 persons. Each eye image has 7 groups and was further divided into training session-1 (containing 3 images) and testing session-2 (containing 4 images). These images were captured with a homemade iris camera which has 8 circularly fitted NIR 850 mn illuminators (Center for Biometrics and Security Research, 2017; Alrifaee *et al.*, 2017). Before making the database accessible to the public, the area of the pupil was replaced with a black circle of constant intensity to compensate for the impact of specular reflection. All the iris images are saved in bmp format at 320*280 resolutions. The capturing framework of the CASIA database is depicted in Figure 2.6.

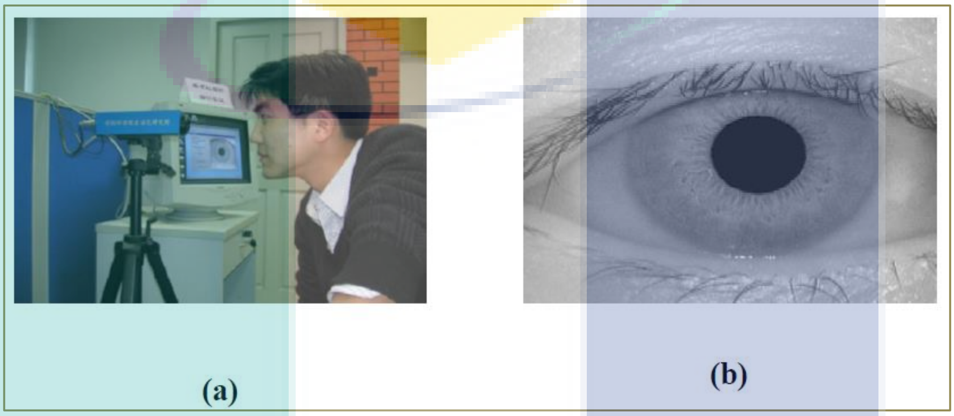


Figure 2.6 CASIA-Iris V1 (a) capturing device, (b) captured image

Source: Biometrics Ideal Test, (2017)

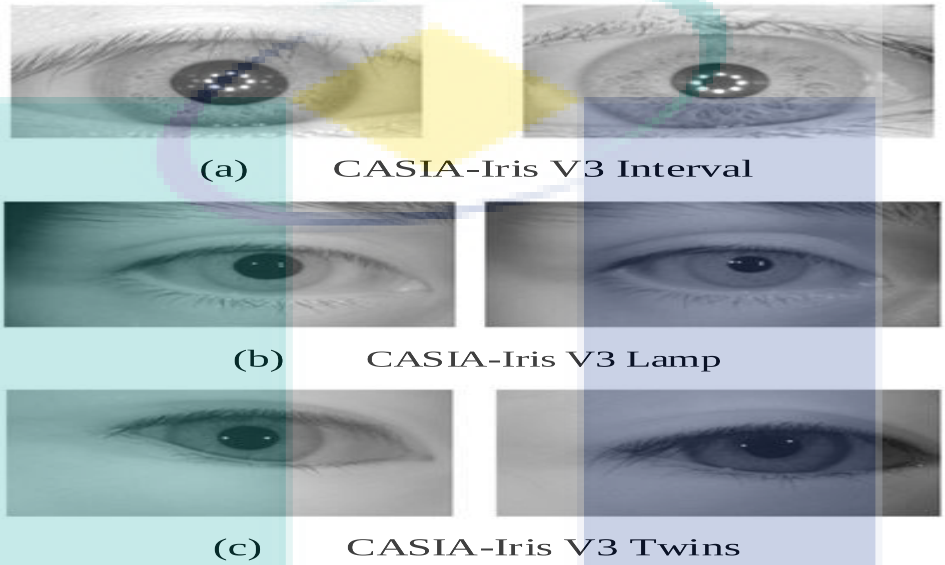
2.4.1.2 CASIA-Iris V2

This database consists of 2400 images which are grouped into 2 subsets as they were captured by 2 devices (Alrifaee *et al.*, 2017). An Irispass-h device built by OKI was used to capture the first 1200 images, while a CASIA-Iris cam developed at the CAS was used to capture the remaining 1200 images. There are 60 different groups of each iris image saved in bmp format at 640*480 resolutions. Figure 2.7 and Figure 2.8 present the imaging devices that had been employed to capture the iris images in CASIA V2 storage.

2.4.1.3 CASIA-Iris V3

This is the 1st CASIA database to introduce noise factors. It comprised of 22,034 iris images which were captured from 700 people. There are 3 subsets in the CASAI-iris V3 database. The Interval data subset contains 2,639 rich-textured images with close proximity and lighted with LED NIR. All the images are saved in jpg format at

320*280 pixel resolution. The Lamp subset, on the other hand, consists of 16,212 iris images which are stored in the database in the jpg format at 640*480 pixel resolution. The Twins data subset contains 3,183 images captured from 100 pairs of twins and saved in jpg format at 640*480 pixel resolution. Some of the images in the CASIA V3 Interval, Lamp, and Twins are shown in Figures 2.9, 2.10, and 2.11, respectively.



(a) CASIA-Iris V3 Interval

(b) CASIA-Iris V3 Lamp

(c) CASIA-Iris V3 Twins

Figure 2.7 Samples from CASIA database

Source: Biometrics Ideal Test, (2017)

2.4.1.4 CASIA-Iris V3

The early iris identification applications dealt with iris images taken with human subject constraints, while the present research is dedicated to toning down these coerces (Center for Biometrics and Security Research, 2017; Alrifaee *et al.*, 2017). The CASIA began the building of a database with new iris images that consisted of images from moving subjects over a distance but with poor quality. The new database was called CASIA-Iris V4. The CASIA-Iris V4 is actually an extension of the CASIA-Iris V3 with additional 3 subsets. The first subset is called CASIA-Iris-Distance, where the iris images were taken over a distance of 3 m while the subject is moving. The framework shown in Figure 2.7 is taken from the CASIA website (Centre for Biometrics and Security Research, 2017; Alrifaee *et al.*, 2017). The database consisted of 2576 images with a resolution of 2352*1728.

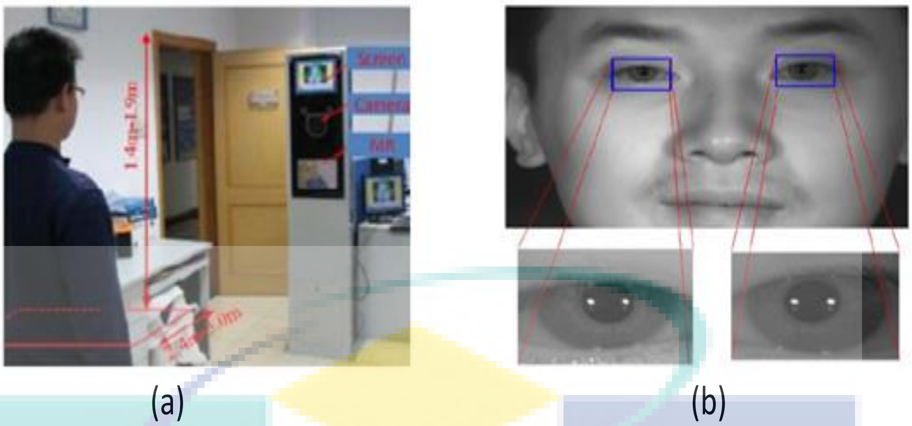


Figure 2.8 CASIA-Iris V4 Distance (a) imaging device, (b) result image

Source: Biometrics Ideal Test, (2017)

The second subset is called CASIA-Iris-Thousand (Figure 2.9). It employed an IKEMB-100 dual camera with a friendly interface. The output images reveal image textures that reflect "what you see is what you get."

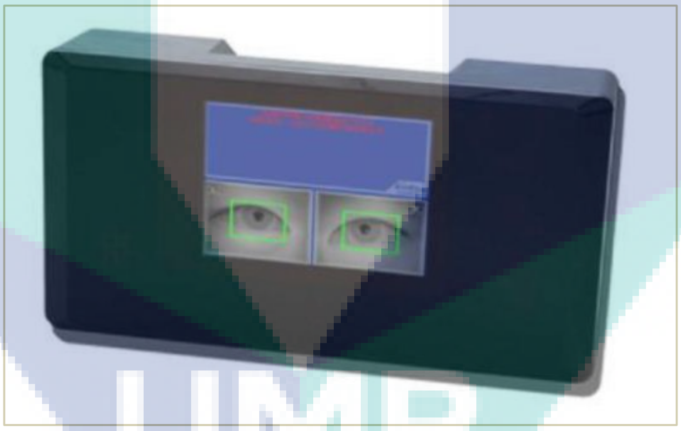


Figure 2.9 CASIA-Iris V4 Thousand imaging device

Source: Biometrics Ideal Test, (2017)

In fact, the bounding box seen around the eye, as illustrated in the previous figure suggests that the subjects should shift their positions to obtain vivid images. This database comprised of 20,000 images at a resolution of 640*480. Lastly, the third set, which is called CASIA-Iris-Syn (Figure 2.10), contains 10,000 images with a resolution of 640*480. The images are derived from a CASIA-Iris V1 in a process suggested by (Center for Biometrics and Security Research, 2017; Alrifaee *et al.,* 2017). This process caused the texture of the iris to be more realistic, thus, enabling it to solve the noise issue caused by rotation, deformation, and motion blur.

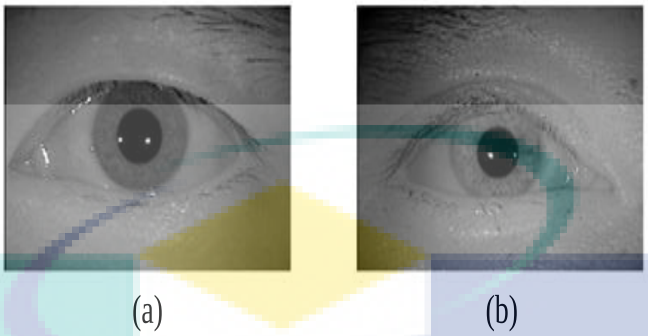


Figure 2.10 CASIA-Iris-Syn

Source: Biometrics Ideal Test, (2017)

2.4.2 BATH Iris Database

This database was developed by Smart Sensor Limited at the University of Bath (BATH Iris Database, 2017; Alrifaee *et al.*, 2017). It is comprised of NIR images, containing 32,000 high-quality iris images taken from 800 mixed ethnic individuals (1600 classes for left and right eyes). The images were saved at a pixel resolution of 1280×960. Some of the sample images contained in the Bath database are shown in Figure 2.11.

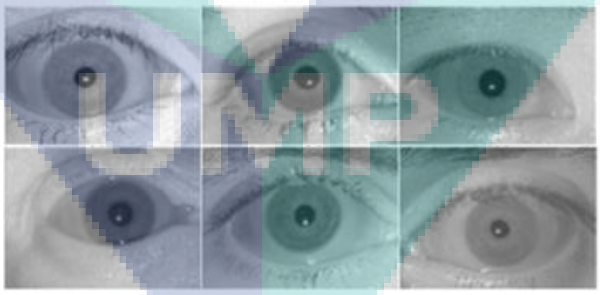


Figure 2.11 Samples from Bath iris database

Source: BATH Iris Database, (2017)

An ISG LW-1.3-S-1394 1.3-megapixel camera positioned over an adjustable base (Figure 2.12) was employed here (SmartSensores, 2014). The camera is equipped with several LEDs to capture 200 images per subject. The best 20 frames were saved in the database. The subjects are required to place their chins on a lever and stood over a short distance from the camera lens. Then, the lens of the camera is adjusted to capture the best and maximum iris textures.

Figure 2.12 Bath database framework

Source: SmartSensores, (2014)

Although some images in the Bath database may be non-ideal, such as those with a diverged look, obstructed by eyelids, and focus blur, most of the images exhibit uniform traits, thus, not ideal for unconstrained identification.

2.4.3 Indian Institute of Technology Kanpur (IITK)

The IITK database contains over 1900 images of the right iris captured from 600 persons (≈ 3 images per person) (DIITK, 2016; Alrifaee *et al.*, 2017). The images were captured using a CCD-based iris camera. Additionally, the IITK database contains few images that were captured in non-ideal conditions. The mood of the images was varied by altering the gaze, source of illumination, eyelid occlusion, etc. The iris images in the IITK database are captured under regulated conditions except for a few samples. The efficiency of the system was tested on the iris image of 20 persons collected with variabilities in the gaze and eyelid occlusion. From Figure 2.12, the iris images belong to the same person with minimal imposition on the subject during image capturing. A comparison of the current Iris Databases is shown in Table 2.1.

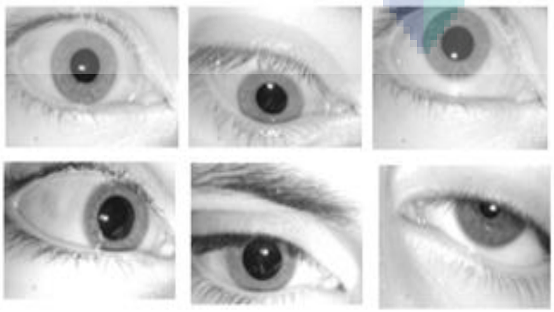


Figure 2.13 Samples iris images from IITK database

Source: Database of Indian Institute of Technology Kanpur, (2016)

Table 2.1 explains the various free iris databases available in public domain to solve the problem of iris biometrics in real world applications. The researcher found that there are two categories of iris image databases that can be used for biometric purposes, cooperative and non-cooperative: In non-cooperative iris images, the user has little or even no active participation in the image capture process, where, the iris images are often captured with more noisy artifacts, such as blur, reflections, occlusions, oblique view-angles, etc., making non-cooperative iris recognition challenging. In this research different iris databases have been used in order to consider all factors such as rotation, noise, scaling, and illumination. Two databases categories were used in this study, the cooperative databases: Academy of Science - Institute of Automation (CASIA) and University of Bath (BATH) and non-cooperative database: Indian Institute of Technology Kanpur (IITK). Besides these standard databases, new four training sets of iris images with different sizes were created randomly from the four used databases using 50 % images from CASIA-IrisV4T, 20% from CASIA-IrisV3I, 20% from BATH, and 10% from IITK database.

Table 2.1 Comparison of Free iris databases

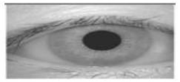
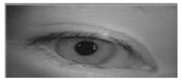
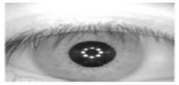
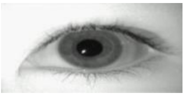
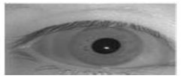
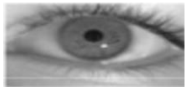
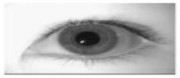
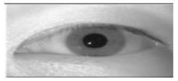
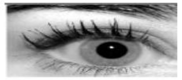
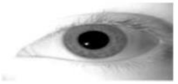
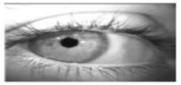
Database	Example image	Database size	Wave length	Varying distance	Camera	Observations
CASIA v1		756	NIR	No	CASIA camera	The previous filling of the pupil regions turns segmentation much easier. High reputable journals now automatically reject papers which experiments were exclusively performed on this ver. of the database.
CASIA v2		2,255	NIR	No	CASIA camera	Subset of the subsequent database version
CASIA v3		22,051	NIR	No	OKI irispass-h	Images captured with two different devices. Contains images with close characteristics to the v 1 ver., with exception of the manual pupil filling.
CASIA v4		54,601	NIR	Yes	OKI's IRISPASS-h	The images were captured in an indoor scenario with nonlinear deformation due to visible light variations. Additionally, it contained several images with poor contrast and heavy blockages.
BATH		16,000	NIR	No	ISG LW 1.3 S 1394	High homogeneous lighting environment. Contains essentially iris obstructions due to eyelids & eyelashes.
IITK		20,420	NIR	No	CCD based iris camera	The images captured under controlled conditions. With change the gaze of the individuals, difference in illumination and occlusion due to eyelids.
MMU 1		450	NIR	No	LG EOU 2200	Noise factors avoided.
MMU 2		995	NIR	No	Panasonic BM ET 100 US	Noise factors avoided.
ICE 1		2,900	NIR	No	LG EOU 2200	Contains off-angle iris images. Intensity values automatically stretched to 171 levels.
ICE 2		75,000	NIR	No	LG EOU 2200	Contains off-angle, partial, rotated and non-iris images and eyes with contact lenses. Intensity val. automatically stretched to 171 levels.

Table 2.1 Continued

WVU		3099	NIR	No	OKI iris-pass h	Contains poor lighting, defocus blur, off angle, and heavy occluded images.
UPOL		384	Visible	No	Sony DXC 950P 3CCD with TOPCON TRC501 A	Completely noise-free images acquired with an optometric framework under high constrained environment.
UBIRIS v1		1877	Visible	No	NIKON E5700	Images captured under heterogenous lighting environments. Several reflections and obstructions can be observed.
UBIRIS v2		11,357	Visible	Yes	Canon EOS 5D	The images were taken under visible lighting while people were walking 4-8m away.

2.5 Data Retrieval

The widespread use of digital technologies and Internet networks, the activity of producing and retrieving data becomes a frequent but still challenging task of retrieving data from large scale databases with satisfactory accuracy and performance rates. Digital image databases are growing very fast and cover a wide variety of application areas. The image retrieval is used in the modern world of digital image process techniques, please refer to Figure 2.14.

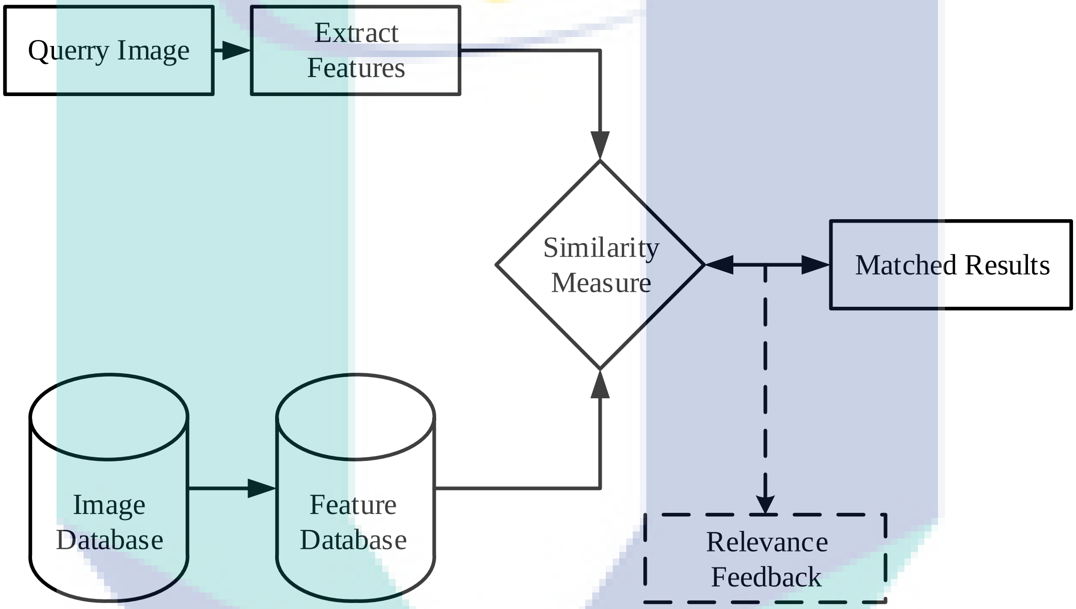


Figure 2.14 Typical retrieval system

Source: Pravin *et al.*, (2012)

That was actually coined by Kato during 1992 (Kakade & Keche, 2017). He has created a system for automatic retrieval of images from a large database based on the features colour and shape. Later, it has been extended into lot and it has been extensively used to get knowledge on retrieving desired image from the large database based on the set of features such as colour, texture, shape and those can be automatically extracted from the images. The features normally can be classified into two kinds. These may be either primitive or semantic. However, the extraction of features should be automatic. The key issue in any kind of image processing techniques is that how to extract useful potential features from the image content. The idea behind this concept is data mining.

The methods of image retrieval are utmost derived from the field of image processing and computer vision and is regarded by some as a subset of that field. However, it becomes unique from others through its emphasis on the retrieval of images with preferred characteristics from a collection of considerable size. The image processing technique includes discrete collections of fields such as image enhancement, image compression, transmission and interpretation. But, image retrieval plays a different role. For example, if an image is compared with a single individual's database record to verify its identity. Here, only two images are taken into account and they are matched. If entire database is verified, then the image retrieval shines over there.

2.5.1 Texture-based Retrieval

Though image retrieval based on texture similarity may not be reliable, it is useful in distinguishing aspects of images with similar colors (Xiaoming *et al.*, 2018). There are several techniques for measuring the similarity of image textures. However, the second-order statistics is the best-established technique which is computed from the query and stored images (Babich, 2012). This technique is used for the calculation of the relative brightness of selected pairs of pixels from an image. These are used to measure the texture values such as the degree of contrast and directionality. Gabor filters and fractals are the alternative techniques for texture analysis.

2.6 Iris Image Retrieval

There are several advantages of the iris-based biometric over other biometric traits. It has been postulated that the iris's content is actually stable during an individual's lifetime. it is stated that every iris is unique and no two individuals have similar iris even if they have the same genetic relationship; in fact, there are differences between the two irises of the same individual. In terms of intricate detail, the iris is rich in unique texture features that can be extracted and used for indexing and matching on a large database (Emad & Norrozila, 2015; Radman *et al.*, 2012). The last decade has witnessed several studies in iris image retrieval as it has gained several attentions globally (Khalaf *et al.*, 2018). The designing of retrieval technique is a challenging task as there is a need to develop retrieval techniques that can achieve a compromise between the 3 main performance measures (the accuracy of the retrieval results must be based on the similarity measures and relevant visual features, the retrieval time must be optimized to ensure an acceptable waiting time, and the processing complexity must have an impact on the resources and computation). There are two main tasks of an image retrieval system, namely similarity measurement, and feature extraction. At the feature extraction phase, the feature vectors of the target image are processed and computed to form the visual index database. Similar images to the search image are retrieved by calculating the predefined similarity measures between the feature vectors of the search image and the features of the virtual index database. The images with close similarity to the search image are then retrieved. It should be noted that image retrieval success is strongly dependent on the selection of an efficient similarity metric, as well as the development of a feature extraction technique that can achieve efficient characteristic differentiation while presenting features with low dimensionality. Some of the commonly used low-level images indexing visual features include colour, texture, shape, or any combination of these features.

2.7 Iris Biometrics: Recognition and Indexing

The popularity of iris biometrics has necessitated the collection of various large-scale real-time databases worldwide. In the UAE, a national border crossing security initiative has been launched. Currently, 27 land, air and sea ports of entry in the UAE are equipped with security system (Mehrotra & Majhi, 2013). India has launched a large-scale project called Aadhaar with the aim of providing a unique identification number to each citizen of the country using iris and fingerprint traits (Khalaf *et al.*, 2018). The UK uses IRIS at some of their automated barriers at certain airports (Patel, 2018). Such systems contain several millions/billions of iris templates. Undoubtedly, a long response time for identification is required. In order to reduce the number of comparisons, methods such as clustering and indexing can be applied. These methods reduce the search space by selecting a proper subset of Iris images from the database prior to matching (Gadde *et al.*, 2010).

Clustering involves partitioning the database into subsets based on certain global characteristics. Indexing, on the other hand, involves assigning an index value to each Iris. Two irises from the same individual may have different index values as both data acquisition and processing can be affected by noise. Thus, a good indexing system should retrieve a correct identity. To minimize the required identification time, the input image can only be compared with the retrieved features. All the current reports on iris data indexing techniques fall into 2 categories which are iris colour-based and iris texture-based indexing (Dey & Samanta, 2014). In the iris texture-based indexing, the index keys are derived from the iris texture while in iris colour-based indexing, the colour of the iris is the source of the index keys. The current recognition and indexing techniques are succinctly described in the proceeding sections.

2.7.1 Iris Recognition Techniques

Among All these biometric identification technique, iris recognition is most prominent technique. Iris recognition systems are gaining interest because it is stable over time. Iris scan has been developing an identification/verification system capable of positively identifying and verifying the identity of individuals. The unique patterns of the human iris are used for overcoming previous shortcomings. The iris indicates the colour part of the human eye. It is a circular membrane of the former face of the ocular sphere. It is pierced with a black hole called the pupil which allows the light penetration to the retina. The iris is used to adapt this light quantity by papillary dilation or constriction.

The iris is a combination of several elements. It is richest distinctive textures of the human. The iris recognition algorithms need to be developed and tested in diverse environment and configurations. Research issues are based on iris localization, nonlinear normalization, occlusion, segmentation, liveness detection and large scale identification. It is required to achieve lowest false rejection rate and fastest composite time for template creation and matching. Over past few years many techniques and algorithms are proposed for effective iris recognition under various constraints. Low constraint IRIS recognition is still an area where considerable work needs to be done and there is huge scope to do.

Alheeti, (2011) proposed an iris recognition technique based on wavelet and equalization techniques that help to identify the power of edge detection operators used for generating the minimum features needed in identifying an iris. In this hybrid technique, 2d Discrete Wavelet Transforms with wavelet masks like Haar and Db2 wavelet transform masks are decomposed, followed by applying edge detection operators like canny, Prewitt, Roberts and Sobel to recognize features, Different type of edge detection makes were applied to recognize the iris features; Canny, Prewitt, Roberts and Sobel. It is clear that the obtained results are approximately similar, but the result of a leads to different details, this is due to high power of canny mask. The proposed method requires high calculation to extract small region textural information

Another method was proposed by Sathish it is a multi-algorithmic iris recognition system, in which iris is segmented by performing the following steps (Sathish *et al.*, 2012). Initially a Gaussian smoothing function and then histogram equalization is applied to improve contrast of iris image. Canny edge detector followed by probabilistic circular Hough Transform is then used to segment the iris. Segmented iris is then normalized using Daugman’s rubber sheet model and then features were extracted by decomposing 2-d Gabor filters on the normalized image. A match score is obtained using Hamming distance matching classifier called Feed forward neural network (NN) algorithm and the results were tested using CASIA database. The Multi-Algorithmic approach combines the features of Hamming distance Classifier and Neural Network Classifier for authentication of iris patterns. The error rate has been reduced with an improvement in feature extraction procedure. It is found that Multiple Classifiers improve the performance of a Biometric authentication with better accuracy, the method did not follow the traditional recognition method and only applicable for small dataset

Rai *et al.*, (2014) have introduced a novel and efficient approach for iris feature extraction and recognition. The authors compared the recognition accuracy with the previous reported approaches for finding better recognition rate than using SVM or Hamming distance alone. The researcher claim for the increase of efficiency, when the researcher used separate feature extraction techniques for SVM and Hamming distance based classifier and proven that the accuracy of the proposed method is good for the CASIA as well as for the Chek image database in term of FAR and FRR, the proposed method Lacks effectiveness for non- cooperative iris image and indistinct in textures.

Frucci *et al.*, (2016) used the approach of watershed transform based Iris Recognition system (WIRE) for noisy images acquired in visible wavelength is presented. Key points of the system are: the colour/illumination correction pre-processing step, which is crucial for darkly pigmented irises whose able do would be dominated by corneal specular reflections; the criteria used for the binarization of the watershed transform, leading to a preliminary segmentation which is refined by taking into account the watershed regions at least partially included in the best iris fitting circle; the introduction of a new cost function to score the circles detected as potentially delimiting limbos and pupil. Iris segmentation has a positive effect as regards the iris code, which results to be more accurately computed, so that also the performance of iris recognition is improved. Accessing the performance of WIRE and compare it with the performance of other available methods, and it uses the two well-known databases. This is UBIRIS 1 and 2. In this proposed method the processing time is not evaluated and does not perform for the traditional non- cooperative databases

Wang, (2018) proposed an improved extreme learning machine (ELM) based iris recognition algorithm with hybrid feature. 2D-Gabor filters and GLCM is employed to generate a multi-granularity hybrid feature vector. 2D-Gabor filter and GLCM feature work for capturing low-intermediate frequency and high frequency texture information, respectively. Finally, the researcher utilize extreme learning machine for iris recognition. Experimental results reveal outperforms the proposed ELM-MGIR algorithm comparing with other mainstream iris recognition algorithms. The proposed approach does not perform for the traditional non- cooperative databases.

2.7.2 Iris-Texture Based Indexing

In the enhanced indexing and clustering schemes, iris texture features are used to determine an index, then, this index is used to reduce the solution space in large iris databases. Some techniques for extracting features have been noted for indexing iris database. One of the most related indexing technique is the technique proposed by (Mehrotra *et al.*, 2009) which is based on energy histogram derived from the iris texture. Normally, the image is stored in the database together with a key during enrolment. Hence, a B-tree data structure is used to save the key value. The acquired iris image is then pre-processed using the Circular Hough transforms (Severo *et al.*, 2018) and the result is a block in a rectangular shape (Saad & George, 2014). In addition, the region of rectangular is further improved to enhance the features of the texture besides making it vary in illumination. Next, the features are extracted using multi-resolution Discrete Cosine Transformation (DCT) (Khayam, 2003). The input image is further divided into non-overlapping 8x8 pixel blocks based on the DCT approach. The block of size 8x8 is rearranged for the transformation of the related coefficients into forms of multiple resolutions. For example, with coefficient D(u, v) for a block, the process of rearranged and kept in Si, where i has been defined by (Mehrotra *et al.*, 2009) as presented in Eq. (2.1) and (2.2)

$$i = \begin{cases} 0 & for\ m = 0 \qquad\qquad 2.1 \\ (m-1)\,3 + \left(\frac{a}{m}\right)2 + \left(\frac{b}{m}\right) & Otherwise \qquad 2.2 \end{cases}$$

Let $m = max\ (a, b)\ for\ \ 2^{a-1} \leq u \leq 2^{a}\ and\ 2^{b-1} \leq v \leq 2^{b}, a$ and b have integer values, while *i* ranges between 1 and 10. Upon rearranging, the coefficients *D*(1, 1), *D*(1, 2), *D*(2, 1), and *D*(2, 2) are kept in sub-bands S_1, S_2, S_3, and S_4, respectively. The arrangement of multiple resolution sub-bands for a block of 8x8 is presented in Figure 2.15.

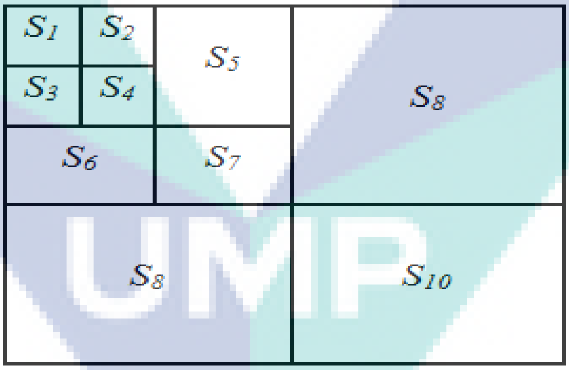


Figure 2.15 Multiresolution rearrangement for 8 × 8 DCT coefficients

Source: Mehrotra, et al., (2009)

After reordering all the DCT blocks, the coefficients from each block belonging to a particular subband (*x*, *y*) are grouped together. Energy value E_i of each subband S_i is obtained by summing up the square of coefficients as, as given below (Mehrotra *et al.*, 2009).

$$E_i = \sum S_i(x, y)^2 \qquad 2.3$$

Note that the sum of square increases the contribution of significant coefficients and suppresses insignificant coefficients. The feature vector consists of different energy values obtained from 10 sub-qbands, as portrayed in Figure 2.16. Other than that, the histogram of energy (H_i) is built for the sub-band (S_i) by utilizing the pictures found in the storage which denotes the sub-band energy distribution. Next, the resultant histogram retrieved from the sub-band is divided into bins. The details of the texture with similar values for energy are grouped into similar bins in order to obtain an accurate match. In fact, images that fall under the bin is illustrated via histogram. As such, the bin might have fixed size or otherwise; however, it remained fixed in this case. Besides, the number of bins can also be employed to develop an indexing universal key, whereby the key of the image should consist of the number of bins corresponding to a sub-band. Hence, the number of bins for a sub-band is re gathered by employing the traversal of Morton order that positions coefficients that are low in proficiency in front of those with higher frequency; for instance, image *I* using Morton order developed the key (3-5-7-8-2-1-4-5-6-7). Likewise, other pictures within the storage should obtain keys as well.

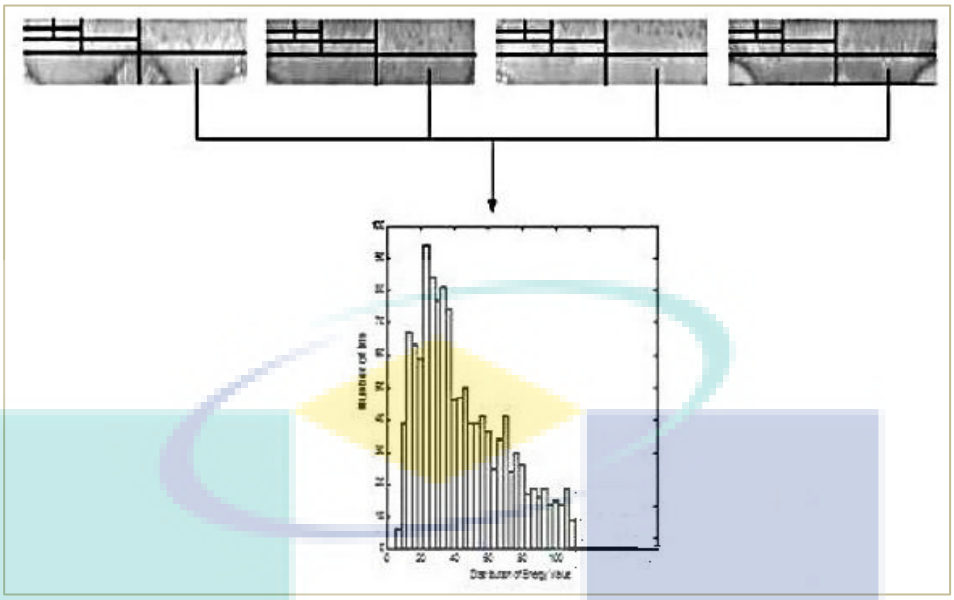


Figure 2.16 Energy Histogram of S10 region

Source: Mehrotra, et al., (2009)

While building a database, B tree is traversed by using a key. Thus, pictures that share similar keys are positioned at similar leaf nodes. Besides, for any query, a key is developed and used to search the tree to the leaf node end. Later, the information stored at the node of the leaf is recovered and after that, crosschecked in order to identify the most accurate match. This approach is ideal for co-operative databases but fails to index non-cooperative irises. Besides, DCT is a known feature extraction technique but it fails to handle variations due to transformations and illumination.

Mehrotra, (2010) suggested two iris database indexing approaches; the first one is the use of DCT coefficients energy histogram to generate a B-tree, while the second one is the indexing of the geometric hashing of SIFT key points. The suggested indexing depends on the detected local key point’s features by the hashing scheme (Wolfson & Rigoutsos, 1997). After the pre-processing and transforming the captured iris image into a fixed normalized size image, the DCT multi-resolution sub-band coding coefficients are used for the extraction of the energy features from the rectangular block. To form an indexing key from the extracted features energy histogram, the key is used to form a B-tree before storing the iris templates at the leaf node with similar texture features. Figure 2.17 depicts the block diagram of the suggested system. (Albuz*et al.,* 1998) proposed searching a large visual database using both content-based image indexing and a retrieval mechanism based on the wavelet coefficients' energy histogram. The approach provides a rapid image retrieval process. A similar approach which considered the energy histogram of reordered DCT coefficients has been suggested by (Wu & Wu, 2002). In the approach, the indexing of the biometrics database is based on the use of the energy histogram of reordered coefficients (Wu & Wu, 2002).

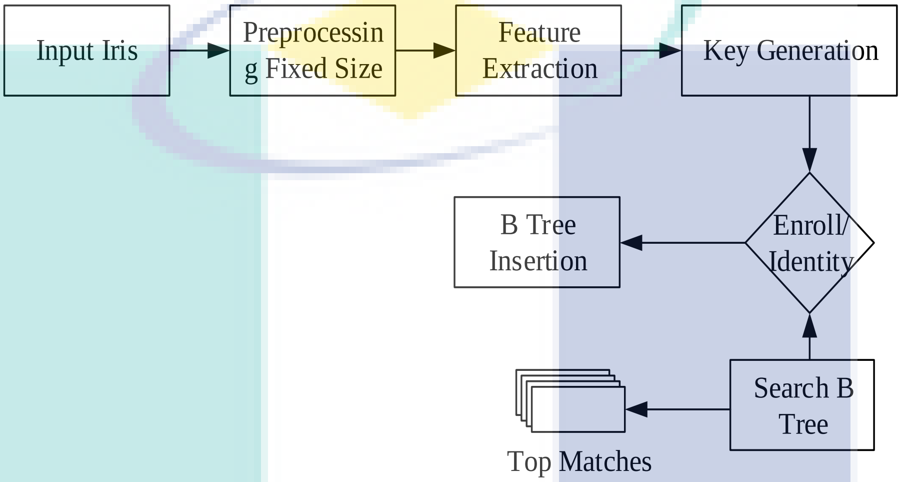


Figure 2.17 Block diagram of DCT based indexing scheme

Source: Mehrotra, et al., (2009)

Nonetheless, indexing based on hashing method does not sort iris template, but otherwise, uses lower-dimensional hashes that are generated directly from the iris data, as proposed by (Rathgeb & Uhl, 2010). In this method, the image was pre-processed by discarding the top and the bottom quarter, say 315˚ to 45˚ and 135˚ to 225˚ of the iris, which consisted of eyelids and eyelashes. After that, the pre-processed texture was utilized to extract a hash based on biometric hash generation (HG) (Uludag *et al.,*, 2004) (Figure 2.18). Once the template obtains a hash, a pointer is generated. The pointer is stored at the related node. If equal hashes were computed for different users, a linked list of pointers would be stored at the respective node. Hence, the node can be reached by short paths. This method overcomes the issue of coarse clustering, whereas the hit rate had been better compared to other methods. However, the limitation of the hashing approach is its scalability and the fact that it functions as purely key generation with no further user-specific secondary data.

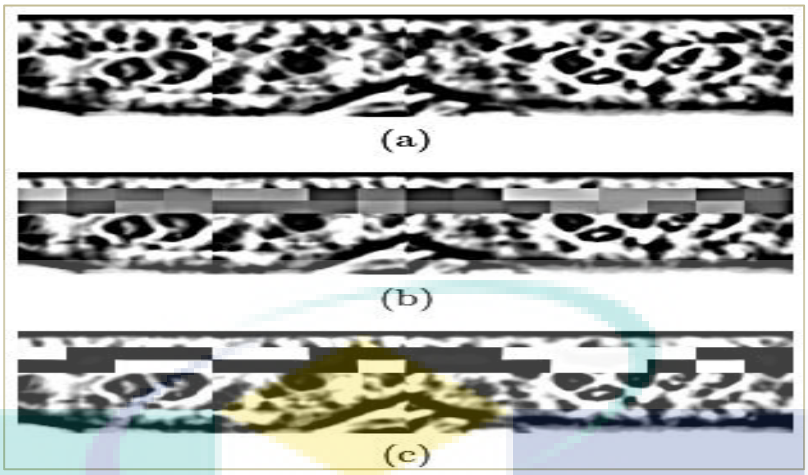


Figure 2.18 The pre-processed steps based on hash generation (HG)

Source: Rathgeb, & Uhl, (2010).

An indexing scheme for iris image database was proposed by (Mehrotraet *et al.,* 2010). The scheme employed for searching large iris database that achieves invariance to illumination, occlusion and similarity transformations. Local descriptors and relative spatial configuration were considered during identification. The local features were extracted from the noise-independent annular iris image using Scale Invariant Feature Transform (SIFT), while the iris database was indexed using the identified key points. Here, a robust geometric hashing scheme was applied. During the iris retrieval, the geometric hashed location of the query iris image was obtained in order to have access to the specific hash table bin. Then, a vote was cast for each entry found and iris images that receive a certain number of votes are the possible candidates. The key point descriptor of all possible candidates was then compared with that of the query iris. Due to the limited searching involved, the query retrieval time is reduced tremendously while improving accuracy. This approach has been applied to UBIRIS, BATH, CASIA and IITK iris databases (Figure 2.19).

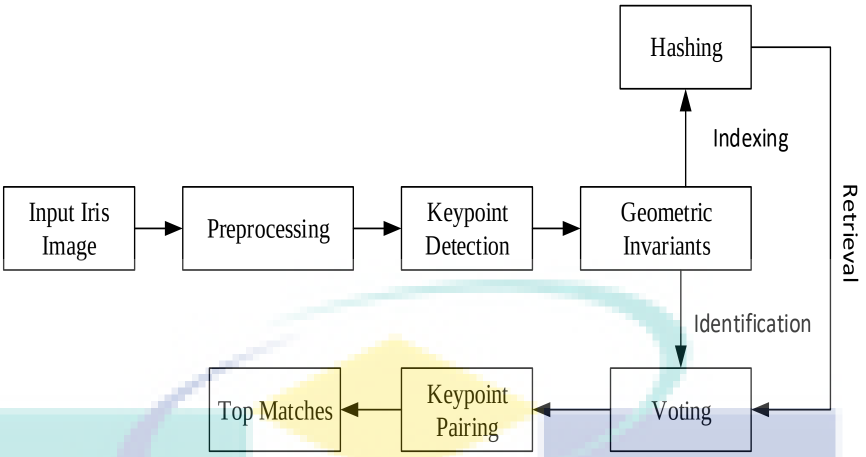


Figure 2.19 Indexing based on hashing of SIFT keypoints approach

Source: Mehrotra *et al.*, (2010)

Dey & Samanta, (2012) suggested an indexing scheme for iris biometric templates retrieval based on the Gabor energy features. These Gabor energy features are estimated from different orientations and scales of the iris texture to generate a 12-dimensional iris image index key. The index key values of all the images are used to create the index space from which the query candidate image is retrieved. After the retrieval, the retrieved images are ranked based on their frequencies. If the query template identity is matched, then it is a hit, otherwise, a miss (Figure 2.20).

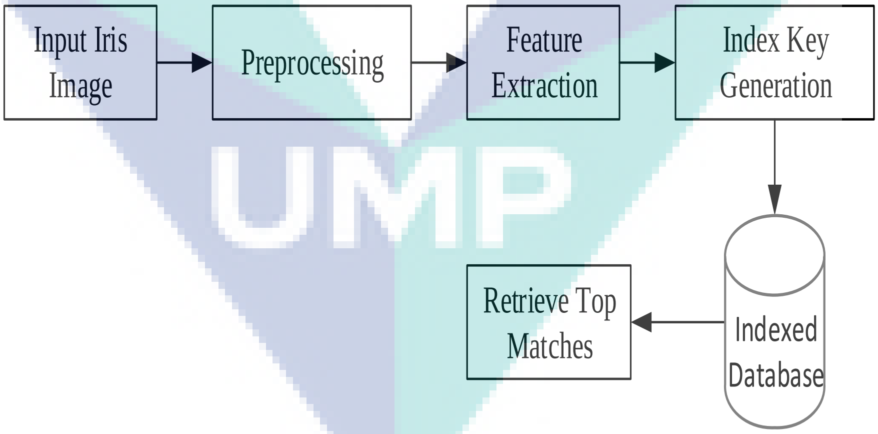


Figure 2.20 Indexing based on based on Gabor energy features

Source: Dey & Samanta, (2012)

Si *et al.*, (2012) attempted one of such approaches. The first step is the detection of the eyelash to promote the iris segmentation accuracy. The researchers stated that the several algorithms misclassify the eyelash as a texture. The proposed scheme which is based on the eyelash directional filters Figure 2.21 and on the notion that eyelashes usually grow in one direction or with a minor is summarized in Figure 2.22, to reduce the loss of the eyelash and to enhance its detection, a connection rule was formed on which eyelashes can be removed based on their positions.

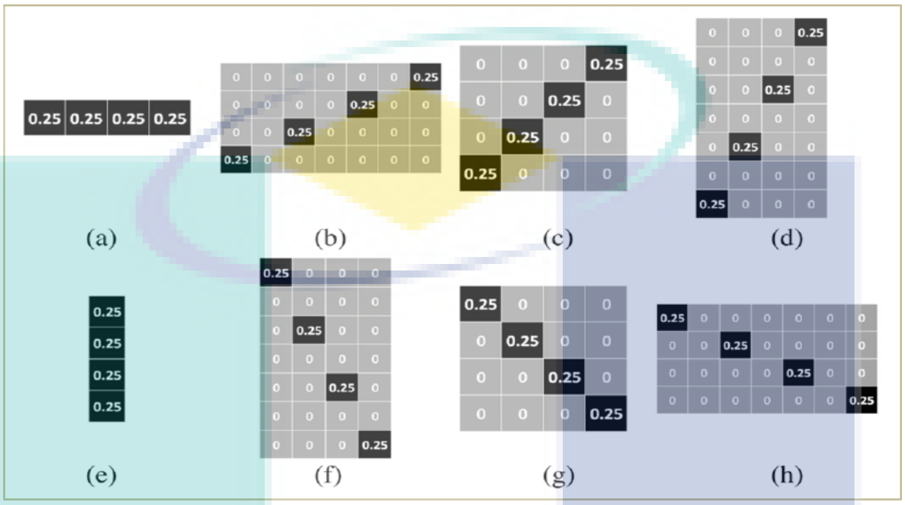


Figure 2.21 Directional filters with eight directions

Source: Si *et al.,* (2012)

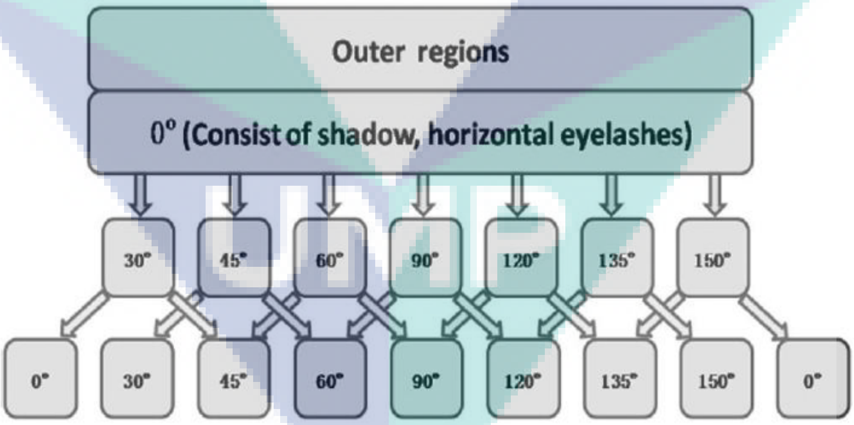


Figure 2.22 Connection rule of the detected pixels (Si, Mei, & Gao, 2012)

The second part is feature extraction which is done based on the multi-scale and multi-orientation data fusion strategy after 2-D Gabor filtering which describes both the scale and direction of the iris texture features. one kind of connection rule can be summarized as Figure 2.23. The local features of the segmented image were extracted by using the scale invariant feature transform (SIFT) method. Upon obtaining the keypoint descriptors from SIFT, a clustering process was performed using the *K*-means method in order to partition the data into m groups. Here, the indexing of key-points was performed based on descriptor property. The k-dimensional tree was built for each cluster centre obtained from the *N* iris images. Therefore, based on these m clusters, m number of k-d trees were obtained and symbolized as ti $(1 \leq i \leq m)$. Those key point descriptors obtained from the probe iris image were clustered into m groups during the retrieval phase. The i^{th} cluster centre was adopted to move it during searching. The *k* nearest neighbour approach was adopted to locate those *p* neighbours from each it that fall within certain radial distance *r* from the probe point in k-dimensional space. Lastly, a total of *p* neighbours from *m* trees were merged and those top *S* matches $(S \subseteq (m\times p))$ that corresponded to the query iris image were obtained. As shown in Figure 2.33, this method has been tested on open databases and its performance is promising.

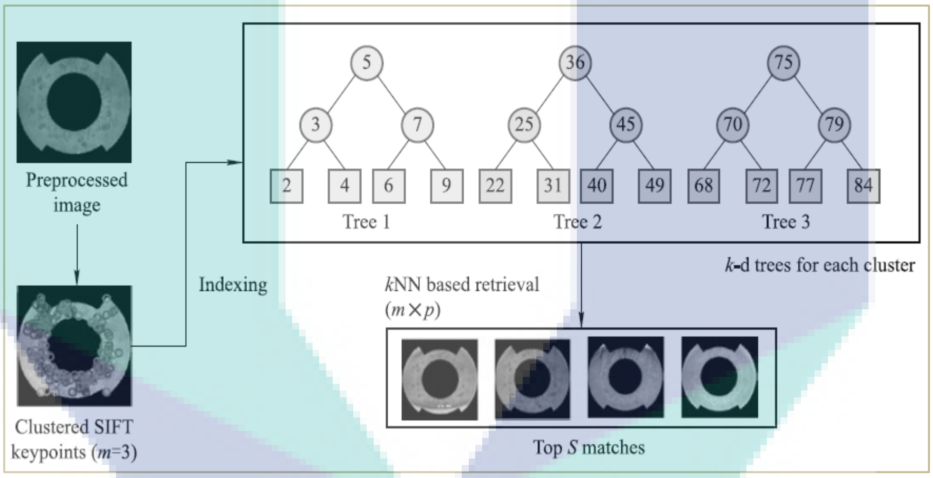


Figure 2.23 Block diagram of the proposed k-d tree based indexing approach

Source: Mehrotra & Majhi, (2013)

The Iris Indexing and Retrieval Model was proposed by (Barbu & Luca, 2015). This model was developed from the concept of HOG-based image feature extraction. In this model, the Histogram of Oriented Gradients (HOG) has been commonly used as a feature descriptor for object-class detection (Dalal & Triggs, 2005). This technique has been proven to be very effective for human identification as well. In this method, the HOG characteristics of color images are determined. A HOG-based feature vector is then built. The K-D-Tree, which is modelled as a binary tree, is used as an indexing solution for image descriptors (Jia *et al.*, 2010) in order to analyse high-dimensional data. The space is recursively partitioned into 2 sub-spaces by the K-D trees. Each nonleaf node generates a splitting hyper plane that divides the space into two (Jia *et al.,* 2010). The points to the left/right of the hyper plane are represented by the left/right subtree of that node. The hyper plane direction is chosen in the following manner. Each node of the tree is associated with one of the K-dimensions, where the hyperplane is perpendicular to the dimensional axis. For example, if the X axis is selected for a particular split, all points residing in the subtree with X values lower than that of the node would appear in the left subtree. Otherwise, they would appear in the right subtree. Here, the hyperplane is prescribed based on the X-value of the point. In this case, its normal vector is parallel to the x-axis. Those points within the HOG-based feature vectors are then added into the K-D tree structure (similar to that of appending an element to a search tree). Therefore, the tree is traversed from the root level, and its direction is dependent on the insertion location of the point. Once the child-node is located, a new point is added. Its location is dependent on the orientation of the splitting plane containing the new node (Figure 2.24).

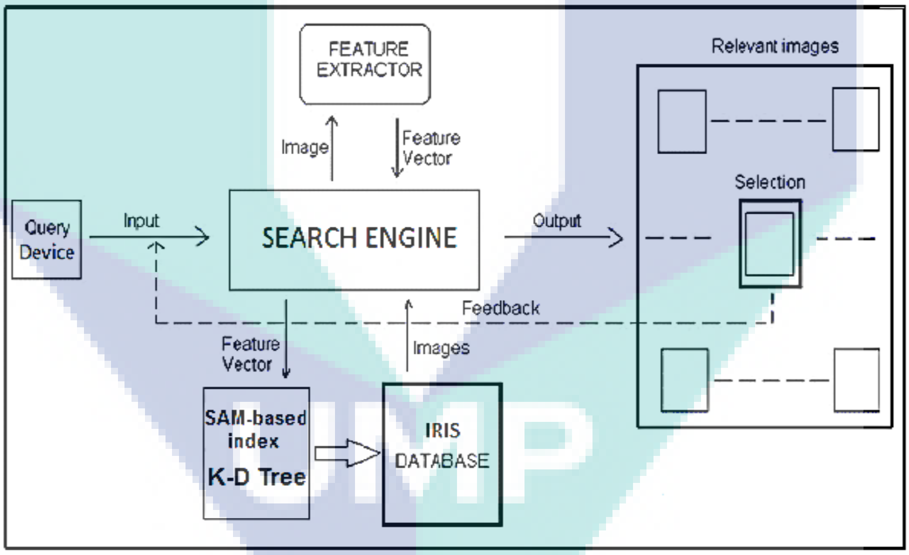


Figure 2.24 Relevance-feedback based indexing and retrieval scheme

Source: Barbu, & Luca, (2015)

Rathgeb *et al.*, (2015) utilized the generic iris recognition system introduced by (Bowyer *et al.,* 2008) for binary feature vectors extraction by conducting a row-wise analysis of the normalized iris textures. It should be noted that iris-codes are typically represented as 2D binary feature vectors which partitioned the binary matrix into blocks of K equal-size (i.e. $w \times h$ bits). Each block is extracted with a Bloom filter in a way that the extracted iris-code B consists of K Bloom filters, i.e. B = {b_1, b_2, . . . , b_K}. The column sequence of the block was transformed sequentially to bits and decimal indexes in order to map a block to B, (B = 1). As an alternative to using several hash functions, H:{0, 1}h → {0, 1}h was utilized, i.e. the sizes of the image set and the inverse image set of H were proportional. The transform was conducted by the mapping of each column c_i ∈ {0, 1}h, i = 1, . . . ,w to its decimal value index as depicted in Figure 2.25. As per Rathgeb *et al.* (2013), the transform is alignment-free to a certain extent, i.e. the generated templates alignment is not important during the comparison. The same number of columns are mapped within certain blocks to the identical indexes of B, i.e. there is a removal of self-propagating errors as a result of iris code misalignment. However, the iris-codes misalignment at the block boundaries seems to distribute several potential matching columns within various blocks, and these blocks are to be mapped to the neighboring B (Rathgeb *et al.*2013). A Summary of the most relevant recognition and indexing methods have been listed in Table 2.2.

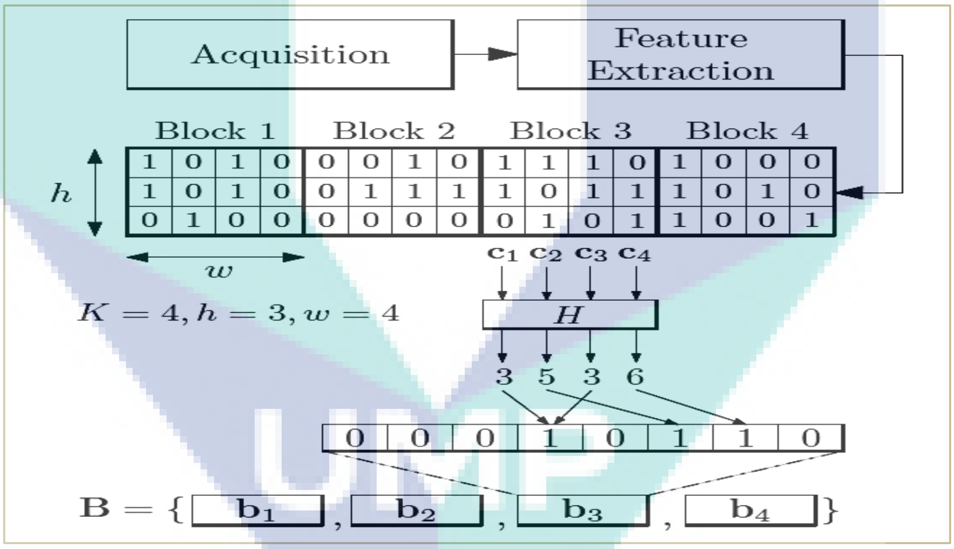


Figure 2.25 Generating a set of Bloom filters from a binary feature vector

Source: Rathgeb *et al.*, (2015).

Table 2.2 A Summary of the most relevant recognition and indexing methods

Method	Author	Database	Summary	Weakness
Hashing indexing	(Rathgeb, & Uhl, 2010)	CASIAv3Interval	Used Hashing to search the templates. The hash function is generated from the iris code and Karnaugh map is constructed for n-bit hashes	*Requires high computational costs and memory as each feature point is severally inserted into the hash table.
Energy Histogram of DCT and Geometric Hashing of SIFT	(Mehrotra, 2010)	UBIRIS, BATH, CASIA, and IITK	Energy Histogram of DCT coefficients and Geometric Hashing of SIFT key points.	*Used one value for indexing iris data which is Energy Histogram of DCT subband.*Computational cost
Indexing based on hashing of SIFT key points	(Mehrotra, Majhi, & Gupta, 2010)	CASIA.1, ICE, WVU(NIR)	Keypoints localization, geometric analysis, hash table construction.	*This indexing method also deals with a large number of key points (approximately 100 key points) with the high dimensional feature vector.
Gabor energy features as indices	(Dey, & Samanta, 2012)	CASIAMMU 2, WVU BATH	Use Gabor energy features in different scale and orientations	*Used only one feature for indexing iris data which is Gabor energy feature from the scales and orientations.
Iris Image Indexing Based on Corner Detection	(Si, Mei, & Gao, 2012)	CASIA V1.0, and IIT Delhi	Used 2D Gabor filters for feature extraction	*Not efficiently reduced the penetration rate (PR) and miss rate (MR).

Table 2.2 Continued

Method	Author	Database	Summary	Weakness
Biometric iris recognition based on hybrid technique	(Alheeti, 2011)	different types of images	based on wavelet and equalization techniques, followed by applying edge detection operators like canny, Prewitt, Roberts and Sobel to recognize features.	Requires high calculation to extract small region textural information
Multi-algorithmic iris recognition	(Sathish, et al., 2012)	The CASIA database	Gaussian smoothing function and histogram equalization, Canny edge detector followed by probabilistic circular Hough Transform.	Did not follow the traditional recognition method and only applicable for small dataset.
Iris recognition using combined support vector machine and Hamming distance approach	(Rai & Yadav 2014)	The CASIA database	The researchers used separate feature extraction techniques for SVM and Hamming distance based classifier and proven that the accuracy.	Lacks effectiveness for non-cooperative iris image and indistinct in textures.
WIRE: Watershed based iris recognition. Pattern Recognition	(Frucci, et al., 2016)	UBIRIS Ver1 and Ver2	Used approach a Watershed transform based Iris Recognition system (WIRE) for noisy images acquired in visible wavelength is presented.	the processing time is not evaluated and does not perform for the traditional non- cooperative databases.
An Improved Iris Recognition Algorithm Based on Hybrid Feature and ELM	(Wang, J., 2018)	CASIA Ver1 and CASIA Ver4 - Interval	Improved extreme learning machine (ELM) based iris recognition algorithm with hybrid feature. 2D-Gabor filters and GLCM.	does not perform for the traditional non- cooperative databases.
Energy histogram of DCT	(Mehrotra, Srinivas, Majhi, & Gupta 2009)	CASIA, BATH, and IITK	Multi-resolution decomposition (DCT). The energy of sub-bands extracted, B-tree indexing.	*Used one value for indexing iris data which is Energy Histogram of DCT subband. *Computational cost

Table 2.2 Continued

Method	Author	Database	Summary	Weakness
Indexing based on k-d trees of SIFT features	(Mehrotra, & Majhi, 2013)	BATH & CASIAV3	Indexing approach is developed using k-d trees of SIFT features, kNN for retrieval.	*This repeated line tracking method is quite time-consuming, especially for large size images.
indexing model based on a HOG-based	(Barbu & Luca, 2015)	UPOL	HOG-based for feature extraction, K-D-Tree for indexing.	Producing high-dimensional feature vectors for indexing.
Towards Bloom Filter-based Indexing	(Rathgeb, Baier & Busch. 2015)	IITD V1	Used Bloom filters for biometric database indexing	*Used high-dim. feature vectors for indexing *Retrieval time relied on database size

In last few decades, good amount of work has been done for recognition but iris based identification is still in its infancy and needs careful attention. An efficient classification, clustering or indexing scheme is required to reduce the search space during identification (Khalaf *et al.*, 2018; Parmar & Degadwala, 2015). There already exist few indexing schemes to partition the biometric database. The existing indexing approaches perform well for cooperative iris databases but fail to achieve desired performance for non-cooperative images vice versa. In other side, most of the existing approaches produce high-dimensional feature vectors for indexing, which is not only increase the computational complexity in indexing, but also increase the logical database size or producing insufficient number of features, this led to misclassification and false indexing. Besides, identification in large biometric systems is sequentially done by comparing the search biometric against all the enrolled templates in the database. This process is computationally costly and increases the systems' response time (Kavati *et al.*, 2015). Based on the current research directions from the literature, investigations have been made in this thesis to present an efficient features extraction and indexing scheme for iris, with an efficient approach to improve response time of the biometric system.

2.8 Clustering Techniques

Clustering is an important part of the whole research steps, shouldering on the responsibility of determining the partitioning and indexing success. The aim of clustering processes is to establish an optimal partitioning for a given set of unknown data; the elements of a dataset are partitioned into clusters that represent their proximate collections. There are many clustering techniques as shown in Figure 2.26. In general, there are two main approaches: hierarchical and partitioned. The former produces a nested series of partitions while the latter produces one partition (Swapna *et al.*, 2016).

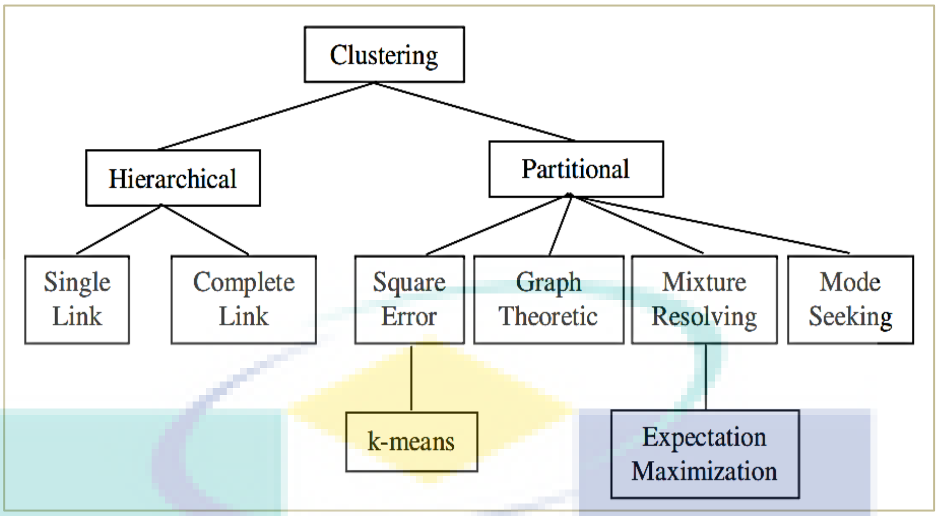


Figure 2.26 Major clustering Approach

Source: Swapna *et al., (*2016)

Clustering is conducted to group similar data objects. Similarity, in general, can be determined via user-specified distance function. For instance, houses can be clustered according to their categories and locations. It is important to ensure low inter-cluster similarity and high intra-cluster similarity. In contrast with classification, predefined classes and class-labelled training examples are not required during clustering and unsupervised learning. In fact, clustering is a form of learning based on observation. Objects are grouped (formation of class) during conceptual clustering. Conventional clustering, however, measures similarity based on the physical distance. Similar to classification, conceptual clustering can be used to discover the appropriate classes and provide descriptions for each class. An example of clustering is laundry, which involves clusters such as permanent press, dry cleaning, washing of brightly-colour clothes, etc. These clusters have some important common attributes. Clustering is straightforward; however, it is difficult to be made. Clustering is generally more dynamic. Some requirements of clustering are:

Scalability: A clustering algorithm must be scalable while handling large dataset.

i. Ability to handle various attributes: An algorithm must be robust enough to handle various data, i.e. interval (numerical), binary, categorical (nominal), and ordinal data, or a hybrid between these data.

ii. Ability to discover clusters of arbitrary shape: A clustering algorithm must be able to detect clusters that exist in various shapes.

iii. Straightforward determination of input parameters: It is desirable to have an algorithm that does not require ample domain knowledge while determining the input parameters. In general, the clustering outcomes are very sensitive to the values of input parameters.

iv. Ability to handle noisy data: A clustering algorithm must be insensitive to missing or erroneous data.

v. Insensitive to the order of input records: A clustering algorithm must be insensitive to the order of input data.

vi. Higher dimensionality: Higher dimensional data objects are highly sparse and skewed. This would complicate the clustering process.

vii. Constraint-based clustering: Clustering might be performed subjected to various constraints. It is challenging to identify groups of data with good clustering behaviour while satisfying specified constraints.

viii. Interpretability and usability: The choice of clustering is highly dependent on semantic interpretation and application.

2.8.1 Partitioning Methods

Given a database consisting of n objects/data tuples, k partitions of data are generated. Here, $k <= n$. The iterative relocation technique is then used to enhance the quality of partitioning by shifting the object to another group. In order to obtain an optimized result, enumeration of all partitions is required. The following heuristic methods are commonly adopted in most applications: (1) the K-medoids algorithm that describes a cluster by using the object located near to the cluster centre; and (2) the K-means algorithm that differentiates each cluster by using the object’s mean value. These methods perform well in identifying spherical-shaped clusters in small to medium-sized databases. While searching for clusters with complex shapes and clustering very large datasets, the enhanced versions of these partitioning-based methods should be developed (Swapna *et al.*, 2016).

2.8.1.1 K-means

This method partitions n observations into k clusters whereby each observation belongs to the cluster with its closest mean. The iterative refinement technique is commonly adopted as it is simple and fast, which is suitable for processing large datasets. Figure 2.27 outlines the K-Means algorithm.

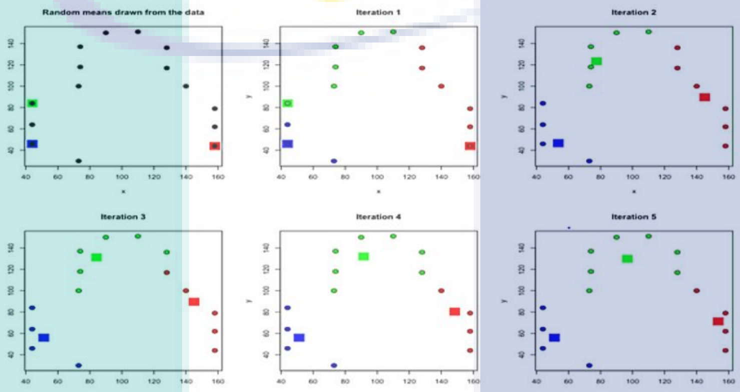


Figure 2.27 K-Means Clustering

Source: Forster *et al.*, (2017)

Here's how the algorithm works (Figure 2.28):

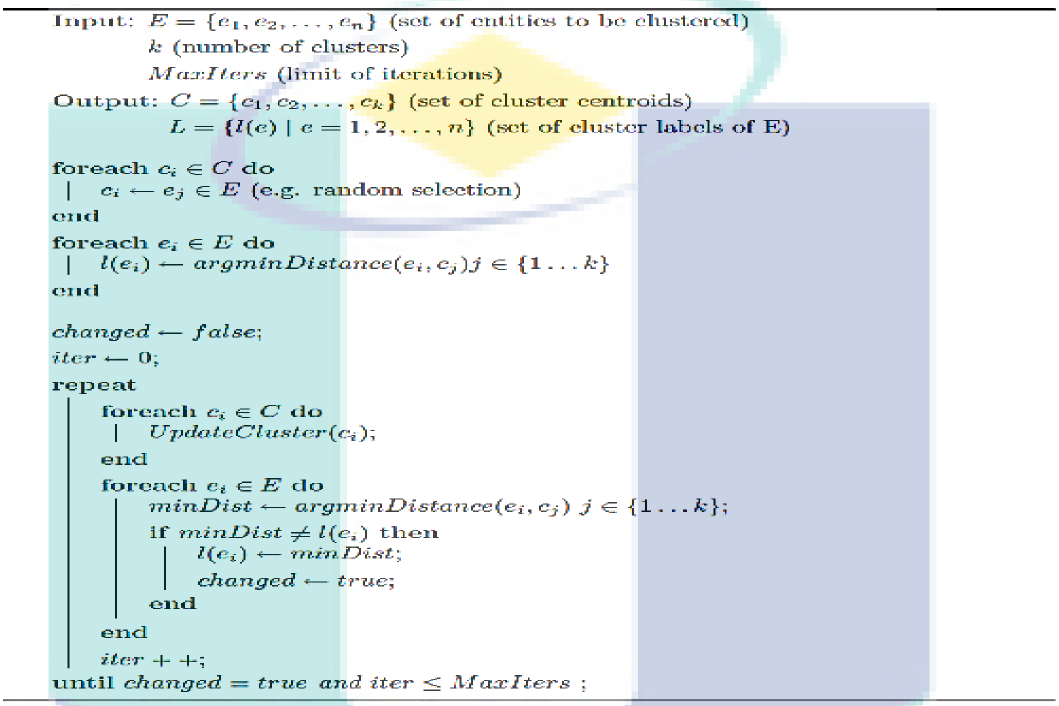
i. *k* points are randomly selected as the centers of the initial cluster ("means").

ii. Upon calculating the Euclidean distance between the points and the cluster center, these points are allocated to the closest cluster.

iii. The center of each cluster is recalculated from the points within the cluster.

iv. Steps 2 and 3 are repeated until convergence.

It should be noted that this algorithm is not deterministic since the resulting clusters rely on the initial random assignments. The main advantages of this algorithm are its simplicity and speed, which allows it to run on large datasets. Its disadvantage is that it does not yield the same result with each run, since the resulting clusters depend on the initial random assignments.



```
Input: $E = \{e_1, e_2, \ldots, e_n\}$ (set of entities to be clustered)
       $k$ (number of clusters)
       $MaxIters$ (limit of iterations)
Output: $C = \{c_1, c_2, \ldots, c_k\}$ (set of cluster centroids)
        $L = \{l(e) \mid e = 1, 2, \ldots, n\}$ (set of cluster labels of E)

foreach $c_i \in C$ do
    $c_i \longleftarrow e_j \in E$ (e.g. random selection)
end
foreach $e_i \in E$ do
    $l(e_i) \longleftarrow argminDistance(e_i, c_j) j \in \{1 \ldots k\}$
end

$changed \longleftarrow false$;
$iter \longleftarrow 0$;
repeat
    foreach $c_i \in C$ do
        $UpdateCluster(c_i)$;
    end
    foreach $e_i \in E$ do
        $minDist \longleftarrow argminDistance(e_i, c_j)\ j \in \{1 \ldots k\}$;
        if $minDist \neq l(e_i)$ then
            $l(e_i) \longleftarrow minDist$;
            $changed \longleftarrow true$;
        end
    end
    $iter + +$;
until $changed = true$ and $iter \leq MaxIters$ ;
```

Figure 2.28 K-means Algorithm

Source: Forster *et al.*, (2017)

2.8.1.2 K-means++ (K-means||)

This method was originally proposed by (Arthur & Vassilvitskii, 2007) to seed the initial centers for K-means. A simple probabilistic method was used to generate the initial centers of points X. It involves the following steps:

- Obtain the first center c[1] from the uniform distribution of points X.

- For k = 2 to K

Sample the k_{th} center $c[k]$ from a multinomial over X where point x has probability θx defined as:

$$\theta_x = \frac{D(x)^2}{\sum_{x^1 \in X} D(x^1)^2} \propto D(x)^2 \qquad 2.4$$

where *D*(*x*) = distance to the nearest existing center

The basic idea of k-means ++ algorithm

This algorithm is mainly used in the selection of seed points by ensuring that the distance between the seed points is as large as possible while eliminating the noise. The procedure is outlined in the following.

i. From the input of the dataset (requires *k* clustering), a point is randomly selected and denoted as the first cluster center.

ii. The distance *D* (*x*) between each point x and the nearest cluster center (i.e. the selected cluster center) is calculated.

iii. A new data point with largest *D*(*x*) is selected as a new cluster center.

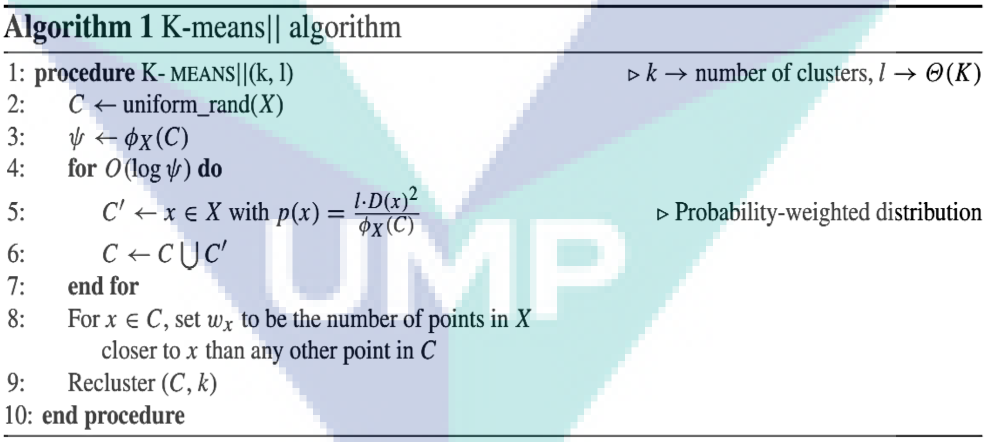
iv. Both steps 2 and 3 are iterated until a total number of *k* clustering centers are obtained.

v. Finally, the initial k clustering center is fed into the standard *K*-means algorithm.

Assuming that set X has n data points, step 2 is executed so that distance data *D (1), D (2), ..., D (n)* are obtained. In order to eliminate noise, one should not directly select the largest value of the element. Instead, the larger value of the element should be selected, followed by choosing its corresponding data points as seed points. In order to choose the larger value of the elements, each element *D (x)* is treated as a line *L (x),* where its length is the value of the element. These lines are connected in the following order: *L (1), L (2), ..., L (n)* in order to form a long line *L*. Here, *L (1), L (2), ..., L (n)* is called *L*'s sub-lines. From the theory of probability, if a point on L is randomly selected, then the point is likely to be lying on the long sub-line. Hence, the data points on this sub-line can be adopted as seed points. Unfortunately, K-means++ is inefficient while handling large datasets. As the dataset grows, the number of classes increases. This condition further complicates the distance calculation procedure.

2.8.1.3 Scalable K-Means++

An efficient method has been proposed by (Bahmani *et al.*, 2012) to overcome the scalability issue of K-means++. Apache Spark has parallelized the K-means++ algorithm called Scalable K-Means++ (or K-means||). It is faster compared to K-means and GMM, thanks to the MapReduce computational model (Bahmani, et al., 2012). The noise management of K-means++ has been modified in order to speed up the computation (less iteration). K-means|| deviates from K-means++ in terms of the initialization method of centroids. K-means|| is initiated by determining the first centroid from the uniformly distributed point set randomly, i.e. $C \leftarrow c^1$. The initial cost of clustering can then be calculated as $\psi = \varphi x(C)$. *Here,* $\varphi x(C) = x \in X\ d2(x,C)$. Each point $x \in X$ with probability $px = l \bullet d2(x, C)\ \varphi x(C)$ is then sampled in log ψ iterations, by inserting the sampled points to C. After each iteration, the number of selected points is expected to be one. A total of l log ψ points are finally grouped in k centers. Figure 2.29 shows the process of Algorithm 1.



Algorithm 1 K-means|| algorithm

```
procedure K- MEANS||(k, l)                              ▷ $k \rightarrow$ number of clusters, $l \rightarrow \Theta(K)$
    $C \leftarrow$ uniform_rand($X$)
    $\psi \leftarrow \phi_X(C)$
    for $O(\log \psi)$ do
        $C' \leftarrow x \in X$ with $p(x) = \frac{l \cdot D(x)^2}{\phi_X(C)}$      ▷ Probability-weighted distribution
        $C \leftarrow C \bigcup C'$
    end for
    For $x \in C$, set $w_x$ to be the number of points in $X$
        closer to $x$ than any other point in $C$
    Recluster $(C, k)$
end procedure
```

Figure 2.29 Scalable K-Means++ (K-means||)

Source: Bahmani, *et al.*, (2012)

The following parameters are required while using the K-means|| algorithm, i.e.:

i. k, which is the number of desired clusters;

ii. max Iterations, which is the maximum number of iterations;

iii. runs, which is the number of runs.

The type of initialization is specified via the parameter called initialization mode. The model option is used to create the initial centers. Similarly, the number of execution steps, the number of initialization steps, and the convergence criterion should be prescribed. The K-means|| operations can be summarized thus: firstly, the data are saved in the slave nodes; next, the algorithmic environment is prepared by executing the master node. Meanwhile, the "for" loop in line 4 of Algorithm 1 is executed by the slave nodes in a parallel manner. Here, a total of 2k points (in average) are sampled for each execution. The probability is proportional to the squared distance measured from the center. Upon exiting the loop, several candidate centers (>*k*) are generated. Each center is weighed by the number of points in the mapped dataset. In order to select only *k* number of centers, a local K-means++ is executed on these centers.

2.9 Clustering problem

Clustering refers to the act of dataset permutation into categories called clusters. The constituents of each cluster are alike but differ cluster-wise. This implies that while there is an intra-cluster similarity, inter-cluster objects differ (Han & Kamber, 2006; Jain, 2010). Over the last 30 years, there has been an increase in the number of disciplines involving image processing which is one of the fields that depends on clustering. It helps the ability of the search engines in making timely and accurate image retrievals from large databases (Bathla *et al.*, 2018). Moreover, it should be noted that clustering is also used in the biology field, especially in cellular processes. Clustering has been successfully employed in the study of various gene aspects. Several other fields have reported the utilization of clustering algorithms, such as in outlier detection, image processing, bioinformatics, document clustering, marketing, and customer analysis (Friedman *et al.*, 2007; Cai *et al.*, 2007; Fan *et al.*, 2009; Kerr *et al.*, 2008; Bsoul & Mohd, 2011; Bsoul *et al.*, 2014; Lee *et al.*, 2013).

2.10 Improvements of K-means Clustering algorithm

K-means algorithm is one of the classical methods of clustering which has been widely studied and used in pattern recognition, data mining, remote sensing image, statistics, and bio-gene information processing because of its fast and easy characteristics. However, this method is sensitive to the initial clustering centre point, and the effect of noise and an isolated point on the clustering effect is relatively large. According to the sensitivity of the K-means algorithm for the initial clustering centre selection, many scholars have improved the initial clustering centre of the K-means algorithm. A method was proposed by Mahmud *et al.*, (2012) to determine the weighted average score of the dataset. The researchers have used a uniform method to find the rank score by averaging the attributes of all data points so that the generated initial centroids tally with the original data distribution. The scores of these data points were then sorted (via a sorting algorithm). A total of k subsets were then generated. Here, k is the number of desired clusters. The centroid with the smallest mean value from each subset was taken as the initial centroid. Still, the desired cluster number should be provided as the input while operating this algorithm.

Another method was proposed by Tidke *et al.,* (2012) where, the ELKI clustering toolkit was used to segregate data mining algorithms and data management tasks for separate evaluation. This strategy, however, is not considered in WEKA, YALE, and GIST. ELKI can handle arbitrary data types, file formats, distance or similarity measures. It has found application in handling real-world datasets such as wine (http://archive.ics.uci.edu/ml/), yeast (http://archive.ics.uci.edu/ml/), and NBA player career statistics. Unfortunately, the associated merging algorithm is highly complex. The method called IKCBD was reported Shunye, (2013) with the adoption of dissimilarity concept. The data is normalized initially. Then, the pre-treated dataset serves as the input to the algorithm. It involves three steps:

i. By calculating x_{maxt}, x_{mint} and x_t of each attribute, the d_m matrix (i.e. dissimilarity matrix) of dataset D is obtained by determining the odd value of each object.

ii. By using d_m and the k values adapted to determine the initial centroids; the Huffman tree is constructed via the Huffman algorithm.

iii. The k clusters are derived by executing the K-means with the initial centroids.

Figure 2.30 shows the enhanced version of K-means algorithm that works based on dissimilarity. Obviously, the computational cost of IKCBD algorithm is dependent on the dataset size (m) and the number of attributes (n). This algorithm should be further tested on a large high-dimensional dataset. Its efficiency can be enhanced by reducing the number of attributes. Also, the value of k should be determined automatically.

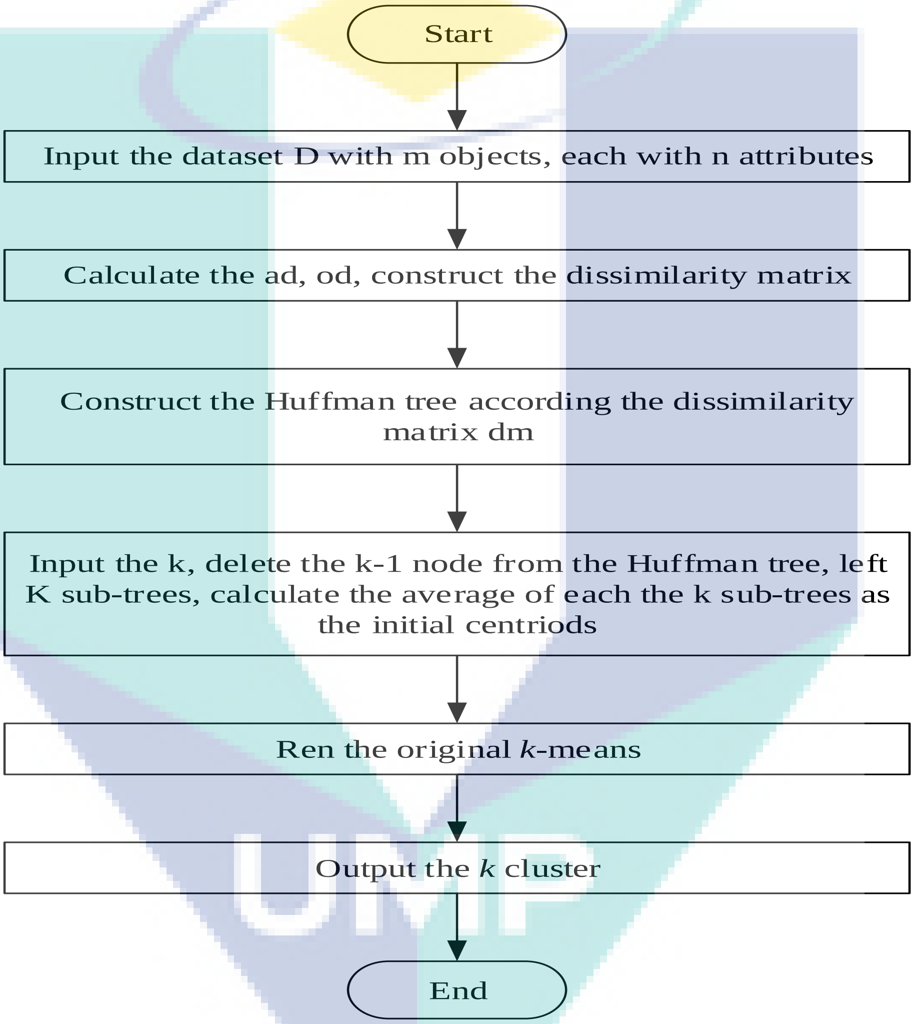

Start
Input the dataset D with m objects, each with n attributes
Calculate the ad, od, construct the dissimilarity matrix
Construct the Huffman tree according the dissimilarity matrix dm
Input the k, delete the k-1 node from the Huffman tree, left K sub-trees, calculate the average of each the k sub-trees as the initial centriods
Ren the original *k*-means
Output the *k* cluster
End

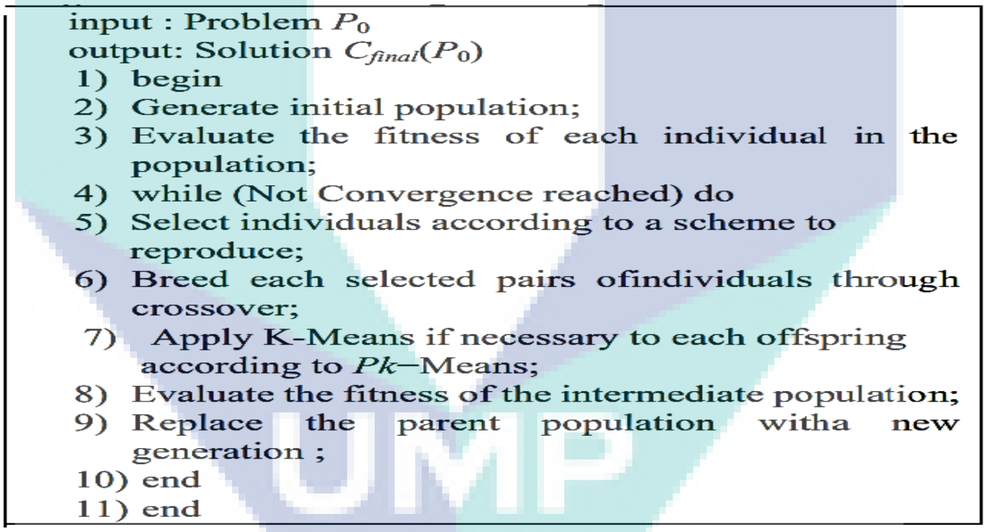
Figure 2.30 k-means clustering algorithm based on dissimilarity (IKCBD)

Source: Tidke *et al.*, (2012)

Yu *et al.*, (2013) proposed a K-means algorithm based on artificial fish swarm (AFS) optimization for solving the problem initial clustering center. The KM-AFS is first initiated by the AFS before recording the optimal food thickness on the bulletin board. Later, the prey action, swarm action, follow action, and random movement action of the fish swarm is respectively stimulated for an intelligent search of the initial clustering centres in parallels, while simultaneously updating the best food thickness and its associated excellent artificial fish with the best food thickness.

The K-means algorithm is then run based on the initial clustering centers generated by the AFS. The K-means algorithm proposed by Chadha & Kumar, (2014) does not depend on the number of clusters as input. Here, two clusters are created initially, which are chosen from those centroids that are farthest apart in the dataset (to ensure dissimilarity). Input: D is a set consisting of n tuples with numeric attributes A_l, A_2, ... , Am. Here, m is the number of attributes. Output: Optimal number of clusters with uniformly distributed n tuples. Method: 1) The attributes of all tuples are summed to located those points that are farthest apart in the dataset; 2) The tuples that containing the minimum and maximum sum are treated as initial centroids; 3) Initial clusters are created after calculating the EDs between the initial centroids and each tuple; 4) The tuple-centroids distance in both initial clusters is determined before recording the minimum non-zero distance (d); 5) The new means for the created partitions in step 3 are determined; 6) The outliers are found by calculating the ED between each tuple and the new cluster centers. No outlier is found if mean<d; 7) The new centroids of the clusters are determined; 8) The Euclidean distance between each outlier and the new cluster centers is calculated to determine the outliers, i.e. mean>d; 9); The set of outliers obtained in step 8 ($B = \{Y_l, Y_2, ... Y_p\}$) is found; 10) The process is continued until $B=\phi$. *: b*). Find the cluster outliers based on the objective function in step 6. a) generate a new cluster for set B using the mean of its members as the centroid. c) If the number of outliers = p, then, i) generate a new cluster using one of the outliers as a component and test each outlier for the objective function. ii) identify the outliers if any exists. d) determine the distance of each outlier from the existing cluster centroids and adjust the outliers in the existing cluster to meet the objective function in step 6. e) *B ={ZI,Z2 .... Zq}*is the new set of outliers, the value of q is dependent on a number of outliers. This algorithm is only applicable to numeric datasets.

Kettani *et al.,* (2015) proposed another algorithm that initiated by setting k = floor ((n) 1/2); where n represents the number of objects in the dataset. This is motivated by fact that this number is in the range of 2 to (n) 1/2 as reported by Pal and Bezdek. The K-means algorithm is deployed with these initial k centroids while the centroid of the least cluster is eliminated. The K-means s then, restarted with the remaining centroids. At every iteration, the maximum CH cluster validity index of the existing partition is saved. The researcher used this index because it is relatively inexpensive to compute, and it generally outperforms other cluster validity indices as reported by Milligan and Cooper as mention in the paper. This process is repeated until k=2. Finally, the algorithm outputs the optimal k and partition corresponding to the maximum value of CH stored so far. However, this algorithm has problems such as the increases in a number of computation's steps because of unnecessary distance calculations; it also needs to enhance the clustering accuracy. A proposed algorithm by Bouhmala *et al.,* (2015) the main idea of this algorithm is to use the genetic search approach to generate new clusters using the famous two-point crossover and then apply the K-Means technique to further improve the quality of the formed clusters in order to speed up the search process. Figure 2.31 shows the steps of the proposed genetic algorithm.



```
input : Problem $P_0$
output: Solution $C_{final}(P_0)$
begin
Generate initial population;
Evaluate the fitness of each individual in the
population;
while (Not Convergence reached) do
Select individuals according to a scheme to
reproduce;
Breed each selected pairs ofindividuals through
crossover;
 Apply K-Means if necessary to each offspring
 according to $Pk$—Means;
Evaluate the fitness of the intermediate population;
Replace the parent population witha new
generation ;
end
end
```

Figure 2.31 The steps of the proposed genetic algorithm

Source: Bouhmala *et al.*, (2015)

The quality of the clustering has always been compared against the solution given and not through the value of the ED cost function which is generally used in literature as it does not represent clustering quality, making it an inappropriate metric for maximizing both intra-cluster homogeneity and inter-cluster heterogeneity. A summary of the related works on clustering frameworks have been listed in Table 2.3.

Table 2.3 A summary of the related works on clustering frameworks

Author	Deployed dataset	Baseline framework	Performed better than
Farnstrom & Lewis (2008)	KDD	Scalable, complete k-means	k-means
Bouras & Tsogkas (2010)	Web	Single, maximum, centroid AHC and k-medoids, k-means++	k-means
Taeho (2009)	News	Single-pass	k-means
Jain (2010)	image	Fuzzy c-means, c-means, k-means, k-means++, k-medoids, Single-pass	k-means
Mohd *et al.*, (2012)	Crimes	Single-pass, k-means	Enhance k-means
Velmurugan & Santhanan (2011)	Geographic map	Fuzzy c-means	k-means, k-medoids
Meila & Heckerman (2013)	image	Expectation-Maximization, hierarchical agglomerative, Fuzzy c-means, c-means, k-means	k-means

One of the major clustering approaches is the centre-based clustering technique, and the k-means algorithm is a representative algorithm based on this technique (Heckerman, 2013). Just a few of such points can have a significant influence on the respective cluster means (Celebi, *et al.*, 2013); hence, the k-means is affected by noise and outlier points despite the fact that it can be applied in many cases with ease. The k-means has been shown as the best clustering algorithm despite its inability to overcome some performance problems (Bouras *et.al* 2010; Jain, 2010). Additionally, its performance is highly sensitive to the choice of the initial centres which may be trapped at the local optima rather than the global optima (Meila and Heckerman, 2013; Raykov & Little, 2016). If the k-means is improperly implemented, there may be empty clusters, a high chance of being trapped in the local optima, and slow convergence (Celebi, 2011). Hence, metaheuristics can be used to overcome these obstacles. Please refer to the Summary of various methods proposed to improve the k-mean algorithm in Table 2.4.

Table 2.4 A Summary of various methods proposed to improve the k-mean algorithm

Method	Author	Summary	Weakness
K-means Clustering algorithm based on weighted average.	Mahmud *et al.*, (2012)	A heuristic method to find better initial centroids and reduce computational time.	Need to provide the desired cluster number as input
Split and merge technique.	Tidke *et al.*, (2012)	The method split and merge datasets for providing a clustering structure that dynamically selects its cluster number and generates an initial clustering result.	The high complexity of merging algorithm and needs to be more efficient.
K-means Clustering Algorithm Based on Dissimilarity.	Shunye, (2013)	The method used Huffman tree to select initial centroids.	*The efficiency of the algorithm needs to be improved to minimize the number of attributes based on PCA. *The k value should present automatic using other methods.
K-Means Clustering Algorithm based on Artificial Fish Swarm.	Yu *et al.*, (2013)	Proposed a K-means algorithm based on artificial fish swarm optimization, which is solved the problem of the initial clustering center.	*The defects due to noisy data and other uncertainties were not solved.
K-Means: A Step Forward for Removal of Dependency on K.	Chadha & Kumar, (2014)	Modified K-means which classifies the input data set into appropriate clusters without taking a number of clusters K as input.	This method is limited to numeric data set.
An automatic clustering algorithm based on K-means.	Kettani *et al.*, (2015)	An alternative parameter-free method for automatic clustering, called AK-means, It is based on successive adequate restarting of K-means.	*Increases in the number of computation's steps because of unnecessary distance calculations *Need to enhance the clustering accuracy.
Genetic Algorithm with K-Means.	Bouhmala *et al.*, (2015)	Used the genetic search approach to generate new clusters using the famous two-point crossover and then apply the K-Means technique.	This framework did not capture the best clusters quality. It is not ideal for homogeneity and heterogeneity maximization within the same clusters and with different clusters, respectively.

The following can be deducted from the analysis of previous works:

i. There are problems related to initial centroids in the partional clusters.
ii. The evaluation of the other process in image clustering like extraction cannot be sufficiently based on the partional clusters as the initial centroids have an effect on the performance decisions.
iii. All the reports approved the use of k-means as a clustering scheme. However, the clustering process and effectiveness of the k-means was not compared with other processes used in other studies.

There is, therefore, a need to select the optimal initial centroids from the optimization frameworks. The next section will focus on meta-heuristics.

2.11 Bio-inspired Optimization Algorithms

The selection of the initial cluster centres and the distribution of data have high impact on the K-means and other algorithms. The latter tends to obtain the local minima rather than the global minimum. The obtained results are highly appreciated in most occasions. This is especially when the chosen initial cluster centres are approximately far apart. This is because it can usually distinguish the main clusters in a given data. Moreover, the K-means' main processing and the dataset's final partitioning quality are both affected by the cluster centroids' initializing process. Thus, the initial points play a key role in the quality of the result. For instance, failure of the K-means algorithm in recognizing the features of the main clusters in certain data is possible if they are close or similar. This failure can also occur particularly if the K-means algorithm is left uncontrolled. Further, associating the K-means algorithm with some optimization procedures is highly necessary. This is in order to be less dependent on a given data and initialization. More importantly, it will enhance the K-means algorithm's performance. Furthermore, it will lead yield good initial clustering centroids and better performance in refining the clustering centroids to find the optimal clustering centres.

The optimization metaheuristics are widely acknowledged as effective method. Two types are worth mentioning here: the single-solution based metaheuristics and the populated-based metaheuristics. The single-solution based metaheuristics are also known as the trajectory methods. The single-solution based metaheuristics start with a single initial solution and move away from it. These metaheuristics describe a trajectory in the search space. Also, some of them can be seen as ‘‘intelligent’’ extensions of local search algorithms. The trajectory methods mainly encompass the simulated annealing method, the tabu-search, the GRASP method, the variable neighbor-hood search, the guided local search, the iterated local search, and their variants. Comparatively, the population-based metaheuristics tackle both a single solution and a set of solutions (i.e. a population). The most widespread population-based approaches are related to the Evolutionary Computation (EC) and the Swarm Intelligence (SI). Darwin’s evolutionary theory has paved the way to the EC algorithms. This is because Darwin’s theory includes the recombination and mutation operators which allow modifying a population of individuals. As for the SI, it mainly aims at creating computational intelligence. This is achieved by exploiting simple analogues of social interaction instead of the individual cognitive abilities (Rana *et al.,* 2011). A Summary of existing Optimization Clustering has been listed in Table 2.5.

Table 2.5 A Summary of existing optimization clustering

Authors	**Type of cluster**	**Optimization framework**	**The Best**
Murthy & Chowdhury (1996)	K-means, GAs	GAs	Gas
Krishna & Narasimha (1999)	k-means, GAs	GAs	k-means+Gas
Kuo *et al.*, (2012)	GAs, PSO	Gas, PSO	Gas+PSO
Runkler (2005)	k-means, ACO	ACO	k-means+ACO
Huang *et al.*, (2013)	k-means, ACO, PSO	ACO, PSO	ACO+PSO
Fun & Ching (2005)	fuzzy C-mean, K-means, PSO	PSO	PSO
Ahmadyfard & Modares (2008)	k-means, PSO	PSO	k-means+PSO
Satapathy *et al.*, (2007)	GA, PSO	GA, PSO	GA +PSO
Yang *et al.*, (2009)	PSO and K-harmonic-means	PSO and K-harmonic-means	PSO+KHM
Zhao *et al.*, (2014)	K-means, PSO	K-means, PSO	K-means+ PSO
Leticia *et al.*, (2014)	k-means, k-majorClust, CHAMELEON, and CLUDIPSO	enhance PSO	enhance PSO
Forsati *et al.*, (2013)	K-means, HS	HS	K-means+ HS
Eskandar *et al.*, (2012)	N/A	Water Cycle, HS, PSO, GAs etc.	Water Cycle
Bose & Mali (2016)	fuzzy C means	ACO, GA, and EM	ABC
Asad *et al.*, (2017)	N/A	ACO, HS, PSO, GAs	ACO
Li *et al.*, (2015)	k-means	DPSO, k-means, CPSO	DPSO
Kumar *et al.*, (2016)	Diverse cluster algorithm	enhance parameter harmony search, harmony search, k-means	enhance parameter harmony search

There are many algorithms that have been proposed in literature to solve the clustering problems. Recently nature-inspired metaheuristic algorithms have been utilized for solving the clustering problem, many stochastic and meta-heuristic algorithms have been recently developed, which can be applied for solving the toughest optimization problem. Nature-inspired metaheuristics have been largely deployed in several fields, including, data mining, computer science, industry, agriculture, (medicine and biology, computer vision, forecasting, scheduling, economy, and engineering (Guo *et al.*, 2009; Christmas, 2011; Chaturvedi, 2008; Zhang *et al.*, 2012; Fox *et al.*, 2007; Cisty, 2010; Rana *et al.*, 2011; Connolly *et al.*, 2012; Akay & Karaboga, 2009; Manoj & Elias, 2012). Metaheuristics are more reliable in handling optimization problems and can achieve adequate solutions within a short time (Karaboga & Basturk, 2008). Additionally, there are two processes in such approaches: the initial solution(s) construction phase and the solution(s) improvement phase. New metaheuristics, such as Firefly algorithm, are usually developed to address the challenges of the existing ones.

Table 2.6 Advantages and Disadvantages of the Main Clustering Methods

Clustering Technique	Examples	Merits	Demerits
Partitioning-based Methods	K-means K-medoids CLARA CLARANS	Simple, relatively scalable and efficient in processing large datasets. Appropriate for datasets with well-separated and spherical clusters.	The distance between objects is not accurate in high-dimensional spaces. It extremely degrades the concepts' performance as almost all pairs of points are considerable distant. Can be stagnated easily at local optima. Not ideal for dense clusters, arbitrary-shapes, and clusters of different sizes. Rely on the user for the determination of the cluster numbers (*k*) in advance. Very sensitive to outliers, noise, and initial centroids.
Hierarchical-based Methods	DIANA AGNES CURE BIRICH CHAMELEON	Flexible with respect to the granularity level. Do not need a pre-knowledge of the number of clusters. Ideal for handling problems that involve point linkages.	Once the combining decision is made, it cannot be corrected. No adequate interpretation for cluster descriptors. Have an inappropriate termination criterion. Needs a high computational time for large and highly dimensional datasets. Has a low performance in high dimensional spaces.

Table 2.6 Continued

Density-based Methods	DBSCAN OPTICS DENCLUE	Discover clusters with different sizes and arbitrary shapes. Resistant to noise and outliers. Good for datasets with low dimensionality.	Very sensitive to input parameters setting. Poor cluster descriptors. Not ideal for highly dimensional datasets. DBSCAN is highly complex in large databases.
Grid-Based Method	STING WaveCluster	Require fast processing time, which does not often depend on the number of data points but mostly on the number of cells in the area. Resistant to outliers and noise. Not affected by the initialisation and input data order. Can detect arbitrarily shaped clusters. Generated output is relatively understandable.	Needs high computational time when faced with highly dimensional datasets. Its complexity is O(n) for low dimensions, but grows exponentially with the dimension. Generates meaningless aggregated information in highly dimensional spaces as most of the data points are mapped into different cells. Hence, grid-based clustering frameworks are only ideal for relatively low-dimensional datasets. The cluster boundaries are either vertical or horizontal, as there are no diagonal boundaries.

Table 2.6 Continued

Subspace Clustering Methods	CLIQUE MAFIA ENCLUS	The computational time grows exponentially with the levelof data dimensionality. The grid-based methods use coarse resolutions for reducing computational complexities and this affects the clustering accuracy.	The computational time grows exponentially with the data dimensionality. The grid-based methods use coarse resolutions for the reduction of computational challenges and this affects the clustering.
Meta-heuristic Clustering Methods	GA PSO ACO HS FA	Tend to move relatively quickly in high-quality solutions, thereby, providing an efficient way of handling large complicated problems. Useful in cases in which traditional clustering methods are stuck at the local optimum.	Cannot guarantee finding an optimal clustering solution. Most of the heuristic methods needed for parameter tuning are dependent on the dataset used. The number of iterations and the initial population can have an influence on the quality of the final solution. Some heuristics are prone to slow convergence rate while some convergence prematurely. Some heuristics have weak local search capability while others have weak global search capability.

As can be seen from Table 2.6 Meta-heuristics clustering methods can be considered as an efficient way to produce more reliable solutions and achieve adequate solutions within a short time (Karaboga & Basturk 2008; Eskandar *et al.*, 2012). Recently, there are new Meta-heuristic algorithms which are inspired from the behaviour of a group of social organisms. These algorithms are called nature inspired algorithm or swarm intelligence algorithms, such as Ant Colony Optimization (ACO), Artificial Bee Colony (ABC), Particle Swarm Optimization (PSO), Bacterial foraging, Bat algorithm (BA), Bee Colony Optimization (BCO), Wolf search, Cat swarm, Cuckoo search, Firefly algorithm (FA), Fish swarm/school, etc. Firefly algorithm (FA) is one of the most promising swarm intelligence algorithm inspired by the flashing behaviour of fireflies (Yang, 2010). Firefly algorithm has two main advantages compared to other algorithms: 1) Automatic segmentation and capable of dealing with multi-quality issues and 2) firefly algorithm takes action based on adsorption and light and is based on swarm intelligence. It could be mentioned that accordingly, such issue can cause automatic segmentation of total population in subgroups with certain average distance and each group can be collected around a local optimum. Moreover, among all of the optimums, the best general optimum could be found. Secondly, the segmentation allows finding all optimums and can be specifically used for nonlinear multi-quality optimization problems (Arora & Singh, 2013). For firefly algorithm, random control is regulated due to iterations, so that convergence can also be accelerated through regulating these parameters. The advantages of being in deal with connectivity, clustering and segmentation and also optimization problems can create an appropriate composition (Farisi *et al.*, 2016).

2.11.1 Firefly Algorithm (FA)

Firefly algorithm is a swarm-based metaheuristic algorithm which was introduced by Yang, (2010), it is one of the well-known swarm-based algorithms which gained popularity within a short time and has different applications. It is easy to understand and implement (Nayak *et al.*, 2016). Firefly algorithm has the following characteristics: it is a natural algorithm. Although the behaviour of a single individual is simple, the collective action can show the remarkable effect, and the algorithm produces the best result. Firefly individuals rely on fluorescence has a strong ability to work together. Firefly individuals tend to move to the optimal location, the brightest individual, and the optimal individuals move randomly to find more locations so that the entire population can have a positive feedback mechanism to ensure that the whole population can find the optimal solution with larger probability. The algorithm is robust. The reason why the firefly algorithm can find the optimal solution quickly is that the feedback mechanism is formed by the information exchange between the individuals in the Firefly, which accelerates the convergence of the algorithm and improves the probability of the best solution. The basic principles of the firefly algorithm are presented in the proceeding sections.

2.11.1.1 The Bionic Principle of Firefly Algorithm

Cambridge scholar Yang Xinshe proposed in 2007 the working principles of fireflies (Nayak *et al.*, 2016). The biological characteristics of information exchange in the fireflies based on their luminous communication with each other inspired the proposed Firefly algorithm (FA). The FA is based on the simulated firefly's luminescence behavior to give some biological significance of firefly light (using only the luminous characteristics with respect to the search area). It depends on the location of a better solution to evolve new solutions (Farisi *et al.*, 2016; Nayak *et al.*, 2016). The bionic principle of the FA is on the use of search space in the simulation of the nature of fireflies. The search and optimization processes are simulated into a firefly attraction and movement process. The function of the standard function is to measure the location of the brighter fireflies. The fittest process can be seen as an iterative search for optimal solvability in the search and optimization processes (Arora & Singh, 2013).

2.11.1.2 Description and Analysis of FA

In the FA, the fireflies attract each other depending on two elements, that is, their brightness and attractiveness. Among them, the fireflies floors with respect to their location and target. Attractiveness is associated with brightness, and the higher the brightness of a firefly, the more it can attract others. If the luminous brightness of all the fireflies in the swarm are the same, the fireflies move randomly. The degree and attractiveness of the emitted fluorescence depend on the propagation medium and brightness of the firefly. The attractiveness decreases as the distance between the fireflies increases (Farisi *et al.*, 2016). To facilitate the description of the FA, 3 idealized rules are usually followed (Arora & Singh, 2013):

i. Fireflies are unisex (male and female) and their mutual attraction is not related to gender.

ii. Attractiveness is proportional to brightness; fireflies with low brightness are attracted to those with high brightness.

iii. The brightness of the firefly is determined by the objective function value of the problem.

The FA is realized based on the following definitions (Tilahun & Ong, 2012): Definition 1: The firefly's fluorescent brightness is:

$$I = I_0 \times e^{-\gamma x r_{ij}} \tag{2.5}$$

$$I \propto f(x_i) \tag{2.6}$$

where, x_i represents the location of the i-th firefly, $f(x_i)$ represents the objective function of the problem. The function value I α $f(x_i)$; I0 is the maximum fluorescent brightness of the firefly, i.e. r=0 where the fluorescence brightness which is related to the objective function value. The objective function value is superior, the degree is brighter. γ is the light intensity absorption coefficient (usually set to constant), r_{ij} is the distance between the fireflies i and j (Tilahun & Ong, 2012):

$$r_{ij} = \|x_i - x_j\| = \sqrt{\sum_{k=1}^{m} \left(x_{i,k} - x_{j,k}\right)^2} \tag{2.7}$$

where m is the data dimension, x_i, k, which represents the k_{th} data component of the firefly i.

Definition 2: The degree of attractiveness of a firefly is:

$$\beta(r) = \beta_0 \times e^{-\gamma x r_{ij}^2} \tag{2.8}$$

where: β_0 is the maximum attraction, that is, the light source at r=0 of the degree of appeal; γ, r_{ij} is the distance between the fireflies i and j.

Definition 3: Firefly i is attracted to the location of firefly j using the update equation:

$$x_{i+1} = x_i + \beta_0 \times e^{-\gamma x r_{ij}^2} \times \left(x_j - x_i\right) + \alpha \varepsilon_i \tag{2.9}$$

where x_i , x_j represents the location of the fireflies i and j; α is the step factor is a constant within [1, 0]; ε_i is subject to Gaussian distribution or evenly distributed random factors, usually used rand - 0.5, where rand is a random distribution of [0,1. $\alpha\varepsilon_i$ is a perturbation term for avoiding premature stagnation in the local optimum.

x_* indicates the location of the brightest fireflies which will be carried out according to Equation 2.6 (Tilahun & Ong, 2012):

$$x_{*+1} = x_* + \alpha\varepsilon_i \qquad 2.10$$

The algorithm is used for optimization processes: The firefly groups are randomly distributed in the solution space because the brightness of each firefly is different from the brightness of the other fireflies. The fireflies with low brightness move towards the brighter ones and the distance depends on the size of the attractiveness. In the location update process, an increase in the disturbance items α × (*rand* - 0.5) will increase the search area to avoid premature local optima stagnation.

The FA has the following characteristics: it is a nature-inspired algorithm. Although the behaviour of a single individual is simple, the collective action can show remarkable effects, and the algorithm produces the best result. The individual fireflies rely on their fluorescence to work together. They tend to move to the optimal location (the brightest individual). The optimal individual moves randomly to find more locations so that the entire population can have a positive feedback mechanism to ensure that the whole population can find the optimal solution with larger probability. The algorithm is robust. The FA can find the optimal solution quickly because it has a feedback mechanism which is formed by the information exchange between the individuals in the swarm. This accelerates the convergence of the algorithm and improves the probability of finding the best solution.

2.12 Common Distance Measures

In most clustering techniques, the selection of a distance measure is an important step as it decides the calculation of the similarity between two elements. The distance measures influence the shape of the clusters as some elements may be close to each other based one distance and farther away based on another. Some of the common distance functions are presented below.

2.12.1 The Euclidean Distance

The Euclidean distance (also called 2-norm distance) is the commonly used distance. Most times, when distance is mentioned, people use to refer to the Euclidean distance (ED). The ED is also simply called distance. It is the best proximity measure to be used when data is dense or continuous. The ED between two points refers to the length of the path that connects those (Moghtadaiee & Dempster, 2015). It is calculated thus:

$$dist = \sqrt{\sum_{k-1}^{n}(p_k - q_k)^2} \qquad 2.11$$

where n = number of attributes, $p_{k,}$ and q_k = kth attributes or data objects p and q.

2.12.2 The Manhattan Distance

The Manhattan distance (MD) is also called taxicab (norm or 1-norm). It is a metric in which the distance between two points is expressed as the sum of the absolute differences of their Cartesian coordinates. Simply, it is the sum of the difference between the x and y coordinates. Assume that the researcher want to find the MD between points A and B, the researcher just have to sum up the absolute mean variation of the x- and y-axes; the researcher have to find how these two points A and B are changing in the x- and y-axes. Mathematically, the MD between two points is measured at right angles along axes (Rajaguru & Prabhakar, 2017). It is calculated thus:

$$d(x, y) = \sqrt[2]{\sum_{i=1}^{p}|x_i - y_i|^2} \qquad 2.12$$

2.13 Transformation Techniques

A number of transformation techniques have been employed in the enhanced system. The following further elaborates the background of each transformation technique.

2.13.1 Discrete Wavelet Transform (DWT)

The DWT is the implementation of the wavelet transforms which uses a set of discrete wavelet scales for functional and numerical analyses (Knitter, 2018). As such, digital filtering techniques are employed to obtain time-scale representation of a digital signal. In the discrete wavelet transform, filters of various cut-off frequencies are utilized to determine the signals at various scales. The discrete wavelet coefficients can be derived by expanding the function *f* (*n*) as a sequence of numbers. The DWT coefficients can be defined through the application of the principle of series expansion as follows:

$$W\varphi(j_0,k) = \frac{1}{\sqrt{M}}\sum_n s(n)\varphi_{j_{0,k}}(n) \qquad 2.13$$

$$W_\psi(j,k) = \frac{1}{\sqrt{M}}\sum_n s(n)\psi_{j,k}(n) \qquad 2.14$$

where $j \geq j_0$ and *s*(n), $\varphi_{j_{0,k}}$ (*n*) and $\psi_{j,k}$ (*n*) are functions of discrete variables *n*=0,1,.....,*M* -1. The DWT coefficients enable the reconstruction of the signal function *f*(*n*) as:

$$s(n) = \frac{1}{\sqrt{M}}\sum_K W\varphi(j_0,k)\varphi_{j0,k}(n) + \sum_{j=0}^{\infty}\sum_k W_\varphi\,(j,k)\psi_{j,k}(n) \qquad 2.15$$

Let j_0=0 and select *M* to be a power of 2($M = 2^j$) , so that the summations are performed over *j*=0,1,.......,*J*-1 and *k*=0,1,2,........2^{j-1}.

The DWT is a better transform it has a better ability in localizing both frequency and time. The calculation of the signal is based on the use of several high pass and low pass filters for the determination of high and low frequencies of the discrete time domain signal, as depicted in Figure 2.32. This is also referred to as the Mallat algorithm or Mallat-tree decomposition (Mallat, 1989).

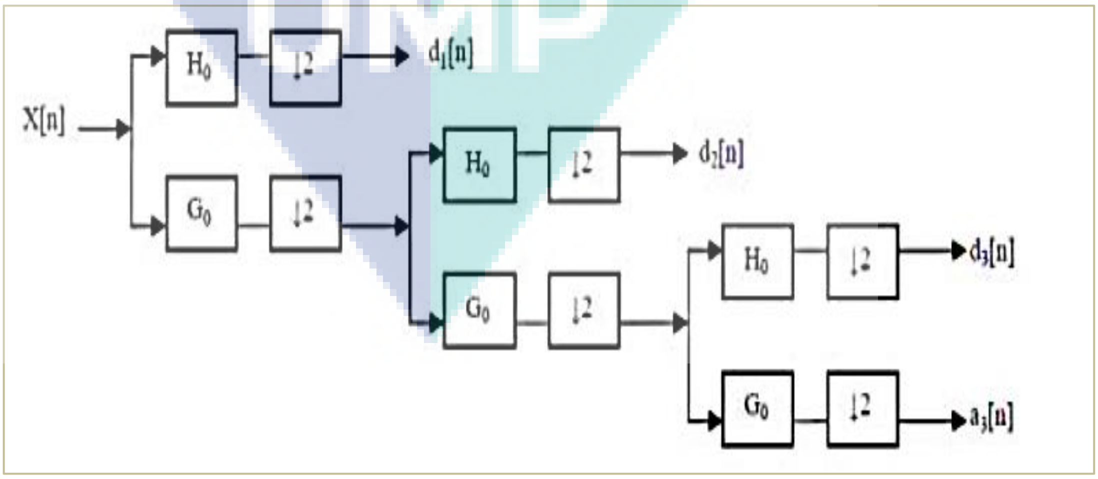


Figure 2.32 Three- level wavelet decomposition tree

Source: Sridhar, (2017).

However, the signal denoted by the sequence $x[n]$ is depicted in Figure 2.32, where n is an integer and G_0 and H_0 represents the high pass and low pass filters. The high pass filter provides a detailed information $d[n]$ at each level while rough approximations an $[n]$ are provided by the low pass filter which is linked to the scaling function. After that, this particular decomposition is repeated to increase the frequency resolution, as well as the approximation coefficients decomposed with high and low pass filters and then, for down-sampling. Besides, this is represented as a binary decomposition tree, with nodes representing a sub-space with some time-frequency localization. The tree is also known as a filter bank.

2.13.2 Discrete Cosine Transform (DCT)

The DCT functions in transforming or converting a signal from a spatial domain into a frequency domain. The DCT has real values and provides a better approximation for a signal with few coefficients. This approach further downsizes the normal equations by disregarding higher frequency DCT coefficients. Moreover, the important structural information is presented in the low-frequency DCT coefficients. Therefore, by separating the high-frequency DCT coefficient and by applying the illumination enhancement in the low–frequency DCT coefficient, the edge information is collected from the satellite images. Besides, the enhanced image is rebuilt by using inverse DCT to obtain a sharper image with good contrast (Mohan & Linda, 2014). Furthermore, in the enhanced technique, initially, the input satellite image, „A‟, is processed by AHE. After that, both images are transformed by DCT into lower and higher frequency DCT coefficients. Later, the correction coefficient for the singular value matrix is calculated as shown below (Mohan & Linda, 2014):

$$\xi = \frac{max(\sum\widehat{D})}{max(\sum D)} \qquad 2.16$$

where $\widehat{D}$ refers to the lower-frequency coefficient singular matrix of the satellite input and output images. The new satellite image (D) is further determined by (Mohan & Linda, 2014):

$$\left(\sum\widehat{D}\right) = \xi\left(\sum D\right) \qquad 2.17$$

$\overline{D}$ is the original images' lower DCT frequency component which is rebuilt using the inverse operation (IDCT) to generate the following equalized image (Mohan & Linda, 2014):

$$\bar{A} = IDCT(\overline{D}) \qquad 2.19$$

Furthermore, this method is measured based on the following significant parameters (Mohan & Linda, 2014):

$$Mean(\mu) = \frac{1}{MN}\sum_{x=1}^{M-1}\sum_{y=1}^{N-1} I(x,y) \qquad 2.20$$

$$standard\ deviation(\sigma) = \sqrt{\frac{1}{MN}\sum_{x=1}^{M-1}\sum_{y=1}^{N-1}\{I(x,y)-\mu\}^2} \qquad 2.21$$

Mean (μ) is the average of all the intensity values. Besides, it represents the images' average brightness, whereby the SD is the deviation of the intensity values around the mean. It also represents the images' average contrast. Here, I (x, y) is the pixels' intensity value (x, y), while (M, N) are the images' dimensions.

DCT Encoding

Two-dimensional DCT is used here and the generalized equation of a 2D (8x8 block) DCT is described thus (Mohan & Linda, 2014):

$$F(u,v) = \frac{C(u)C(v)}{4}\sum_{x=0}^{7}\sum_{y=0}^{7} f(x,y)cos\left[\frac{(2x+1)u\pi}{16}\right]cos\left[\frac{(2y+1)v\pi}{16}\right] \qquad 2.22$$

$$where: C(u), C(v) = \frac{1}{\sqrt{2}}\ for\ u,v = 0 \qquad 2.23$$

$$F(u,v) = \frac{C(u)C(v)}{4}\sum_{x=0}^{7}\sum_{y=0}^{7} f(x,y)cos\left[\frac{(2x+1)u\pi}{16}\right]cos\left[\frac{(2y+1)v\pi}{16}\right]$$

$$otherwise\ C(u), C(v) = 1$$

2.13.3 Singular Value Decomposition (SVD)

The SVD is a numerical method which can be used in numerical analysis to diagonalize matrices. The SVD has been applied in several applications. Its main properties which distinguished it from the other image processing applications include the good stability of the the singular values (SVs) of an image, i.e., there is no significant change in the SVs when a small perturbation is added to an image; as well as the use of the SVs to represent the intrinsic algebraic properties of an image. The SVD method in linear algebra has been employed to solve numerous mathematical problems. Other than that, several approaches are possible in SVD. Each SV reflects the image layer luminance, whereas the corresponding pair of SVs represents the image geometry. Singular values portray the intrinsic algebraic properties of an image. From the aspect of image processing, images are considered as a matrix with no negative scalar entries. The SVD of an image A with size m × m is given by A = U ∑V T, where U and V represent orthogonal matrices, and S = diag(λi) represents a diagonal matrix of SVs λi, i = 1, . . . , m arranged in a descending order. Furthermore, the U columns are the left SVs while the V columns are the right SVs for image A. This process is also called a Singular Value Decomposition (SVD) of A which can be expressed thus (Madhesiya, & Ahmed, 2013):

$$A = USV^T = \sum_{i=1}^{r} \lambda_i u_i v_i^T \qquad 2.24$$

where r represents the rank of A, while ui and vi are the left and right SVs, respectively. Additionally, it is necessary to know that each SV is a luminance representation of the image while the associated SV pairs reflect the intrinsic geometric properties of the images. It has been shown that slight SV variations do not affect the visual perception of the cover image. This motivates watermarking embedment through few SV alterations in the segmented images. The synthesis of this section is illustrated as the previous work do not take into account the powerful of combine between three of them to extract sufficient numbers and most relevant of local features from iris images, which the main problem related to this process is the weakness of each one , so in other meaning, combine three of the transformation methods will fill the gaps related to each one.

2.14 Measures of Performances

The association of two passwords can be known when two alphanumeric aspects get to match perfectly. Nonetheless, biometrics does not crosscheck similar data. In fact, variances should exist when crosschecking data, for instance, age factor. Therefore, the sets of data may vary even for the same person, which is known as intra-class variations. Nevertheless, when differences are spotted for two persons, this condition is termed as inter-class variations (Mehrotra, 2010). However, if two biometric datasets are crosschecked to detect intra-class variations, the results are termed ‘similarity scores/genuine scores’, whereas ‘imposter scores’ reflects inter-class similarity. Furthermore, results that exceed an onset value (τ) is known as false acceptance, while a genuine score below τ is known as false rejection. These measures are further elaborated in the following sub-sections:

2.14.1 Penetration rate (PR)

Penetration rate (PR) refers to the ratio of user samples that can be retrieved from a database based on the presented query template. Let P_c represent the number of rightly probed samples among *P* number of probes, and R_i be the number of retrieved samples from a database of *N* size for an i_{th} probe; then, PR is defined as (Dey & Samanta, 2012; Kavati *et al.*,2017):

$$PR = \frac{1}{M}\sum_{i=1}^{M}\frac{C_i}{N} \qquad 2.25$$

Where C_i represents the individual set of the i^{th} test image, *N* represents the total number of images in the dataset, and *M* represents the number of images to be scanned. A good indexing scheme will achieve a high hit rate (low BM) and a low PR.

2.14.2 Rate of Bin Miss (BM)

When a verification probability is put into a bin, it is called a case of bin error. Such case occurs because the data are not cross-checked with an accurate bin, thus resulting in a match failure. Such a mistake takes place because of wrongly placed biometric data in an incorrect bin while verifying.

2.14.3 Complexity Analysis

Time and space complexities are both referred to as retrieval complexity. In computation science, time complexity of a framework is a function which quantitatively describes the run time of the framework. This is a function of the length of the string representing the value of the input to the algorithm. Time complexity is often expressed by Big O notation (Svagerka, 2018), excluding the lower term and the first term of the function. In this way, time complexity can be called asymptotic, and it examines the case when the input value approaches infinity. Algorithms with more executions take more times than those with fewer executions. The number of executions in an algorithm is referred to as the frequency of a statement or the frequency of time denoted as $T(n)$.

2.14.3.1 Time Complexity

For time-frequency, the time spent on the implementation of an algorithm cannot be theoretically calculated, but the researcher cannot and do not need to test each algorithm on the machine just know which algorithm spends a longer time and which one takes a shorter time. Algorithms take time which is proportional to the number of executions in the algorithm. Algorithms with more executions take more times than those with fewer executions. The number of executions in an algorithm is referred to as the frequency of a statement or the frequency of time denoted as $T(n)$. Time complexity in the frequency of time mentioned earlier n is called the size of the problem, and when n is constantly changing, the time-frequency $T(n)$ will also change. But sometimes, the researcher wants to know what's going on when it changes. To this end, the time complexity concept was introduced. Generally, the frequency of repetition of a basic algorithmic operation depends on the problem size n, denoted by $T(n)$. in the presence of an auxiliary function $f(n)$, as n tends toward infinity, and $T(N) / f(n)$ becomes a constant which is not equal to 0, then, $f(n)$ is said to be in the same magnitude order of $T(n)$, where $T(n) = O(f(n))$, and $O(f(n))$ = algorithmic progressive time complexity, called time complexity.

Although the frequency may vary, time complexity may remain the same. For instance, $T(n) = n^2 + 3n + 4$ and $T(\mathrm{n}) = 4n^2 + 2n + 1$. There is a difference in their frequency but their time complexity remained $O(n^2)$. There are lots of time complexity expressions; for instance, logarithmic order $\mathrm{O}(\log^2 n)$, linear log order $O(n\log^2 n)$, k times order $O(n^{\mathrm{k}})$, square order $O(n^2)$, linear order $O(n)$, order of order $O(n)$, Cubic order $O(n^3)$, ..., exponential order $O(2n)$. As the problem size increases, the time complexity also increases while the algorithmic implementation efficiency reduces. Worst-case complexity and average time complexity: Worst-case time complexity is also referred to as worst-case complexity. Generally, it is not stated that the time complexity of the discussion is the worst case of time complexity. This is because worst-case time complexity is the upper bound of the algorithm's runtime on any input instance, which ensures that the runtime of the algorithm is not extended.

Time complexity in the worst case is *T* (*n*) = *O* (*n*), meaning that the algorithmic runtime cannot be more than 0 (*n*) for any input instance. The average time complexity is the expected run time of the algorithm in the case where all possible input instances occur with equal probability. Algorithms with time complexity of exponential order *O*(2*n*) are slow and cannot be applied when the value of n is slightly larger. The retrieval times of the current algorithms were analyzed to compare. For the indexing method, templates were retrieved for a given search template before performing either searching or matching on the retrieved data. The indexing system is assessed for efficiency in terms of the systems' retrieval time. The retrieval times of the currently used indexing systems were analyzed. In the indexing process, a serial of templates was retrieved for any given query template.

Dey & Samanta, (2012) retrieved similar templates which correspond to 12 indexes key values of the search template, thereby, requiring 12-time comparisons. Again, an algorithmic retrieval time does not rely on the database sizes. The retrieval time complexities were kept at *O*(1). With the 12 comparisons used in this method, the execution time is difficult to be assessed because of the numerous comparison operations. The method uses k-means clustering to store iris data into multiple clusters. For each input sample, the closest matching cluster is determined by the method, indicating that the input template must be matched with all the cluster centers to establish the matching clusters. With this technique, a minimum number of comparisons is required to find the number of clusters; the retrieval time complexities in this method is kept at *O*(1). The key idea of the half-searching is to assume that the length of the array is *N*, then the second is *N* / 2 ,…until the end of the second to the last. This is the worst case as there is a need to find the number of points at every time. The number of times is the number of basic statements; so, the researcher can set the number of times *x*, *N* * (1/2) *x* = 1; then *x* = log *n*. Therefore, in order to overcome the disadvantages of linear searching in the existing methods, the half-searching method was used to match values inside each group in the enhanced method, with the time complexity always kept at *O*(log *n*). Compared to the time complexity *O*(*n*) in linear searching, the half-searching method can improve the computation efficiency.

2.14.3.2 Space Complexity

A programs' spatial complexity is the size of the memory required to run a program. Using the program's spatial complexity, one can pre-estimate the memory needed to run a program. In addition to the need for storage space and the instructions, constants, variables, and input data used by the store itself, a program requires some work units that operate on the data and stores some ancillary space for the information needed to actually compute. The storage space required for program execution consists of (1) Fixed part: The size of this part of the space and the number of input/output data has nothing to do with the value. It mainly includes space occupied by instruction space (ie code space), data space (constant, simple variable) and so on . This part belongs to the static space; (2) Variable space: This part of the space includes the dynamic allocation of space and the space required for the recursive stack. The space size of this part is related to the algorithm. The storage space required for an algorithm is represented by $f\ (n)$. $S\ (n) = O\ (f\ (n))$, where n is the size of the problem, and $S\ (n)$ represents the spatial complexity (Dey & Samanta, 2012).

2.15 Summary

This chapter presents a literature review of the enhanced clustering and indexing methods including the searching and retrieving of iris biometric system. The free iris databases available in public domain have been reviewed. It was observed that there are two categories of iris image databases that can be used for biometric purposes, which are cooperative and non-cooperative. Furthermore, review of the most relevant recognition and indexing methods is presented in details with proper clarification on their drawbacks. Also, the related works on clustering frameworks have been reviewed with explanations on the advantages and disadvantages of the k-means clustering algorithm. This was followed by a summary of various methods proposed to improve the k-mean algorithm and to determine the issues relating to the existing methods. On the other hand, the researcher have reviewed the existing optimization clustering methods with the advantages and disadvantages of the main clustering methods, where the Meta-huristice or bio-inspired optimization has been explained in details and the most promising swarm intelligence algorithm was observed to be the Firefly algorithm (FA). In addition, the distance measurements are well reviewed and presented in details. Furthermore, the necessary background with some other concepts and definitions needed and may be used from time to time have been covered and explained in this chapter.

CHAPTER 3

METHODOLOGY

3.1 Overview

This chapter discussed the approaches used in this study, starting with the description of the thesis groundwork, then, to the explanation of the pre-processing and features extraction approaches. The second part presented the improved clustering method (WKIFA) for partitioning and clustering. The last part explained the efficient searching approach based on parallelization, half-searching, and b-tree. The framework of the research is shown in Figure 3.1. The chapter is structured as follows: The section 3.2 explained the groundwork in this thesis, consisting of the problem identification and iris images. Moreover, section 3.3 explained the pre-processing step for the extraction of the iris portion (ROI) from the image. This section also presented the enhanced features extraction approach and the analysis of used methods. Also, section 3.4 presented the details of the improved clustering method (WKIFA) for achieving a high accurate clustering and indexing. Furthermore, section 3.5 explained the efficient searching approaches based on parallelization, b-tree, and half-searching. The section 3.6 finally presented a summary of the chapter.

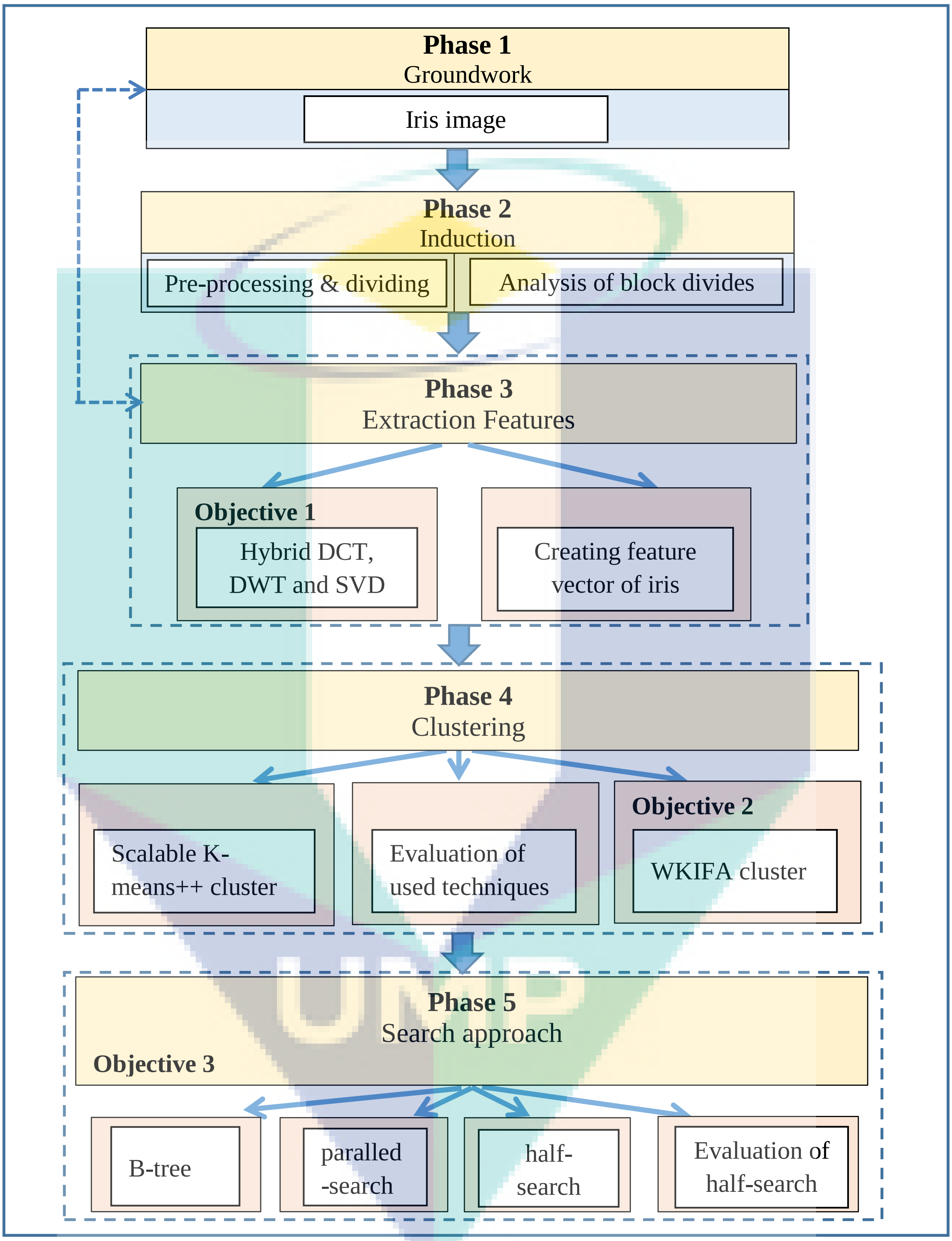


Figure 3.1 The research framework

3.2 Groundwork Phase

The groundwork in this thesis is related to the problem of clustering by reviewing the literature on each problem. This literature review aimed to detect the methods and algorithms that have weaknesses related to iris image clustering and to show the characteristics of iris as an image. In order to improve the potential for developing a theory, this section consists of problem identification and iris image case. This phase begins with identifying the most relevant works. In particular, it concentrates on understanding the challenges in developing effective image clustering to support retrieving iris image. This was achieved by reviewing the state of the art of quality iris image clustering (as presented in Chapter II) in order to identify the strengths and limitations of the current approaches which motivated us to use high performance iris image clustering to analyse the quantity of these clusters of image groups.

3.3 Induction Phase

The reduction of the solution space requires a form of database clustering and indexing (Dey *et al.,* 2014). The response time of the query search depends on the similarity of the template to the query input, and not on the number of templates contained in the database. The database should thus, be logically partitioned in order to minimize the solution space (Claramunt *et al.*, 2015). In the current methods, the each sub-bands' energy value serves as a local feature for the partitioning of the database images into logical groups (Dey & Samanta, 2012; Mehrotra, 2010). Such methods often result in false partitioning and indexing because of the representation of an image with an insufficient number of features. On the other hand, indexing with high dimensional feature vectors can increase the computational difficulties during the retrieval process (Jayaraman *et al.*, 2012; Barbu & Luca, 2015).

In this method, a sufficient number of the most relevant features were extracted from each local region of iris images. The method of extracting the improved feature provided in this study depends on the hybrid Discrete Cosine Transformation (DCT), Discrete Wavelet Transform (DWT) and Singular Value Decomposition (SVD) for the extraction of the local features of iris images for the purpose of partitioning and classifying them into groups using the scalable K-means++ algorithm. The experiments have been done with an Intel Core i7-2600K Quad-Core Processor, 3.4 Ghz, 8 MB Cache, 8.0-GB Memory) with a MATLAB R2014a implementation. Different iris databases have been used in order to consider all factors such as rotation, noise, scaling, and illumination. Two databases categories were used in this study, they are the cooperative databases: Academy of Science - Institute of Automation (CASIA) and University of Bath (BATH) and non-cooperative database: Indian Institute of Technology Kanpur (IITK) (detailed in chapter II). Sample images from these databases are depicted in Figure 3.2.

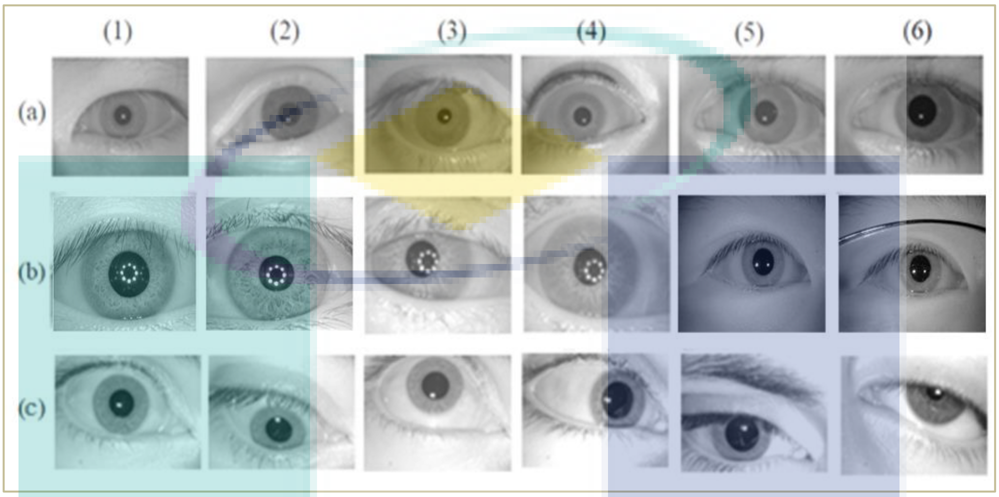


Figure 3.2 Sample of iris images sourced from a) BATH database, b) CASIA database, c) IITK database

Besides these standard databases, new four training sets of iris images with different sizes were created randomly from the four used databases using 50 % images from CASIA-IrisV4T, 20% from CASIA-IrisV3I, 20% from BATH, and 10% from IITK database. The results obtained with these 4 datasets are shown in Appendix A Table A.1. The global block diagram of the enhanced method is shown in Figure 3.3. The next subsections will explain the details of each process.

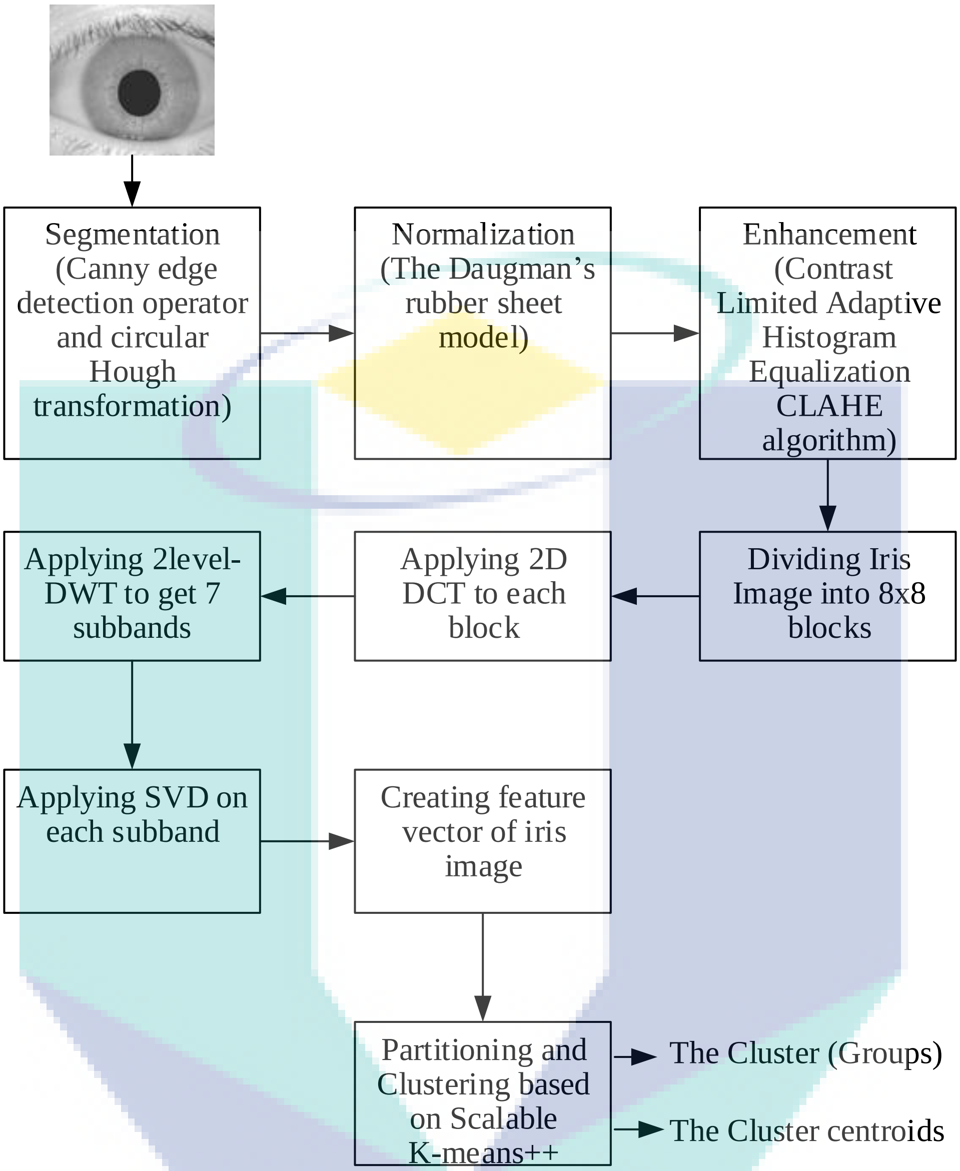

Segmentation (Canny edge detection operator and circular Hough transformation)
Normalization (The Daugman’s rubber sheet model)
Enhancement (Contrast Limited Adaptive Histogram Equalization CLAHE algorithm)
Dividing Iris Image into 8x8 blocks
Applying 2D DCT to each block
Applying 2level-DWT to get 7 subbands
Applying SVD on each subband
Creating feature vector of iris image
Partitioning and Clustering based on Scalable K-means++
The Cluster (Groups)
The Cluster centroids

Figure 3.3 The global block diagram of extracting and clustering local features of iris image

3.3.1 Iris Image Pre-processing

The first step of iris recognition is isolate the actual iris region, this process is called a pre-processing algorithm which is applied to provide the determination of the boundary of iris within the eye image and then extracts the iris portion which is the region of interest (ROI), at the same time remove some of the un-useful parts (e.g., eyelid, pupil, etc.), in order to facilitate its processing. The pre-processing involves stages like iris segmentation, iris normalization and enhancement Figures 3.3 and 3.4 show the iris image pre-processing stages where: (a) Original eye image from CASIA

Database, (b) Applying Canny edge detector, (c) Applying Hough transform, (d) Isolated iris region, (e) Normalized iris.

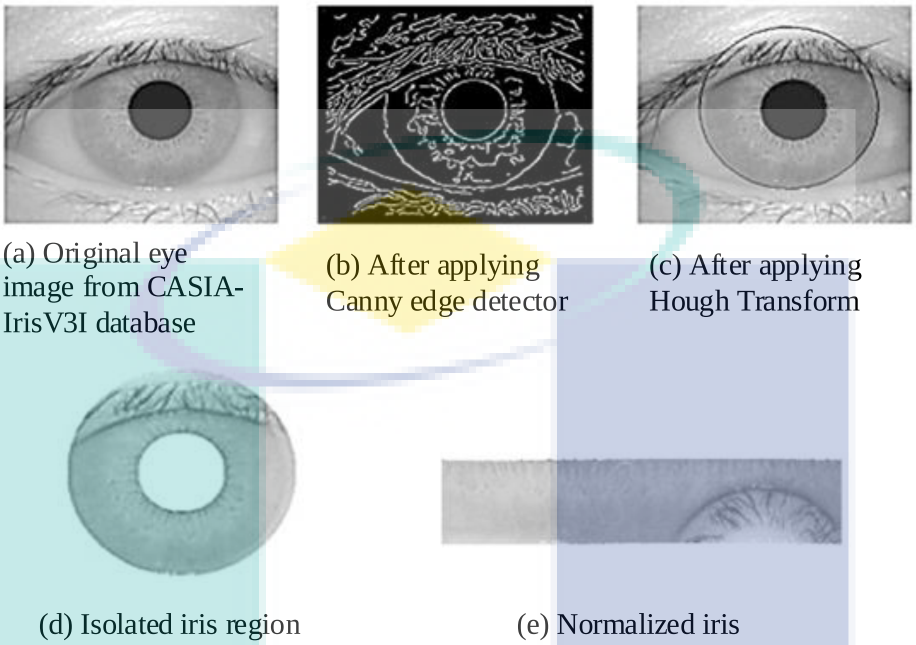

(a) Original eye image from CASIA-IrisV3I database
(b) After applying Canny edge detector
(c) After applying Hough Transform
(d) Isolated iris region
(e) Normalized iris

Figure 3.4 Iris image pre-processing steps

3.3.1.1 Segmentation

For iris segmentation, it has been applied Canny edge detection operator and circular Hough transformation (Kaur & Pathania, 2016). Canny Edge Detection technique is very well known and a popular edge detection algorithm. Though several edge detection techniques such as Sobel, Canny, Prewitt etc. are available, it was observed that Canny edge detection technique is able to extract most of the iris texture from the enhanced image. It is less likely than the others to be fooled by noise and more likely to detect true weak edges. The Canny method finds edges by looking for local maxima of the gradient of image. The gradient is calculated using the derivative of a Gaussian filter. The method uses two thresholds, it is firstly reduce the noise of the image this process called image smoothing, after that calculating edge strength and edge direction then obtain thin edges across the image, finally obtain only the valid edges in an image by invoking threshold with hysteretic. After applying canny edge detection technique Circular Hough is applied to the canny edge detected image for detecting the pupil (inner boundary) and the sclera (outer boundary), it also used for detecting the upper and the lower eyelids. This method is very efficient for the task of finding the iris from an image. Because it works even when noise is present in the image and performs well even when a large amount of the circle is hidden.

3.3.1.2 Normalization

After successfully segmented the image, The Daugman's rubber sheet model (Hanaa & Farag, 2015) is used for normalization. The Daugman's algorithm is being widely used and it tested in different conditions and it has a zero failure rate. Wildes et al. was tested with 520 images and also has no failure. Anwar, (2016) tested with 6000 iris images and it has a recognition rate with 98.4% (Anwar, 2016). The Daugman's algorithm converted the Iris disk to a rectangular region with prefixed size. Points within the iris region are remapped to a pair of polar coordinates (r, θ) where r is on the interval [0, 1] and θ is angle [0,2π]. The remapping of the iris region from Cartesian coordinates (x, y) to polar coordinates is carried out as in (Anwar, 2016). Normalized iris image after applying Daugman's rubber sheet model on segmented iris image is shown in Figure 3.4(e).

3.3.1.3 Image Enhancement

The image is enhanced before feature extraction by using Contrast Limited Adaptive Histogram Equalization CLAHE algorithm (Martíne & Ramos, 2014; Hanaa & Farag, 2015), the average accuracy of image recognition where in the pre-processing stage using the CLAHE technique is 98.06%. This result is better than image recognition performance using other techniques at the pre-processing stage (Nugroho, 2018). The enhancement process is important in order to reduce the noise and to solve the problems of low contrast and illumination of image, and then the image will be ready to divided into 8x8 blocks. the next section will explain the dived image.

3.3.1.4 Dividing Iris Image

This step involves the division of the iris image I into 8x8 blocks as shown in Figure 3.5. The image-textural features are better analysed in detailed sub-images (Kekre *et al.*, 2011). The enhanced method treats each block individually when extracting the relevant local features. The analysis of the small block makes the transformation methods interesting and useful for image processing (Cheng & Yu, 2014).

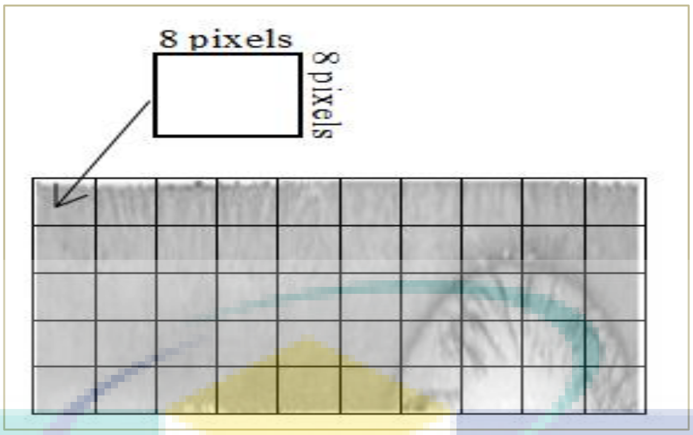


Figure 3.5 Divided image into a block of 8 x 8 pixels

The 8x8 block size is commonly used in the coding of images and videos because it shows a high rate of signal confusion and can be implemented with ease on most computing frameworks (Tyagi, 2015). The *n*×*n* block discrete transform of an image *f(x,y)* of size $S_u \times S_v$ results in a 2*D* array of an equal size as the query image. An analysis and comparison have been done in the next section on the use of different image block sizes, such as 8 x 8, 16 x 16, 32 x 32, 64 x 64, and 128 x 128 on the performance of image retrieval and classification schemes.

I. Analysis of Block Divides: Image local features are small patches which have been extracted from an original image. The extraction of local features from an original image requires the partitioning of the image into small blocks with individually computed features. These features are termed local features because they have a small block size. These local features can achieve good outcomes in several clustering problems (Song et al., 2015). There are certain interesting properties of these local features which suit them for image recognition, such as being inherently strong against translation. They also have interesting properties which are ideal for image retrieval. For image retrieving and clustering, feature vectors of various sizes such as 8 x 8, 16 x 16, etc are generated (Figure 3.6).

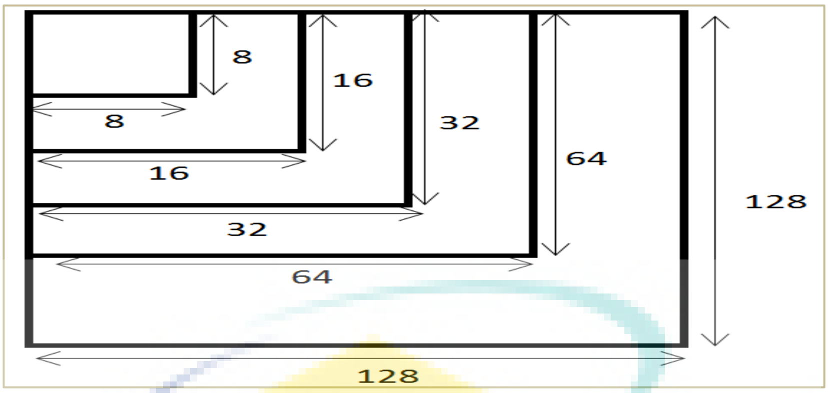


Figure 3.6 Selection of varying size portion from feature

Results of this approach by (Kekre *et al.*, 2011) have been evaluated regarding the optimal block size for a retrieving and classification system, where the presented algorithm is worked over database of 1000 images spread over 10 different classes. The Euclidean distance is used as similarity measure. A threshold value is set to determine to which category the query image belongs to. The well-known Discrete Cosine Transform (DCT) to generate the feature vectors for the purpose of search and retrieval of database images. According to the approach RGB images are converted first into gray level image. Then for spatial localization, the DCT transformation is used. Each image is resized to *N***N* size. DCT is applied on the image to generate a feature vector as shown in Figure 3.7.

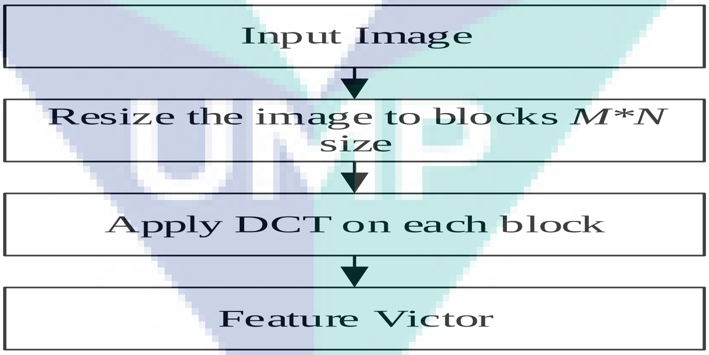


Figure 3.7 Flowchart for feature extraction

a) Algorithm for image clustering

i. Generate the feature vectors of the search image as shown in Figure 3.7.

ii. Compare the generated feature vector to the feature vectors of all the images contained in the database. Use the Euclidean distance (ED) measure to check the closeness of the search image and the images in the database.

iii. Sort the Euclidean distance (ED) values in an increasing order to find the first 100 near matches with the search image.

iv. Calculate the closest matches of the query image for all 10 categories.

v. Set a threshold value for the determination of the category of each search image.

vi. Show the category of the search image.

b) Result analysis

The approach is tested on the image database of 1000 variable size images collected from Corel Collection (Wang *et al.,* 2001) and Caltech-256 dataset (Griffin *et al.,* 2007) These images are arranged in 10 semantic groups. Results showed the average identification rate for 10 types of query images for different image size, as well as the average accuracy of Image Classification. Table 3.1 shows the results of one type of images, a Dinosaur image. The results of this approach showed that the best average identification has been achieved when using images of size 8 x 8 compared to images of other sizes. The analysis of results of the average identification rate for different image types and sizes is shown in Table 3.2. The final result showed that the highest rate was achieved when using images of 8 x 8 size compared to other sizes. From the result analysis, it was found that the best iris image size before local feature extraction was 8 x 8.

Table 3.1 The average rate of identifying 10 Dinosaur query images for all the categories using different feature vector sizes (Kekre et al., 2011)

NO	Feature Victor Size	Rainbow	Mountains	Horse	Rose	Elephant	Dinosaur	Bus	Sea	Coins	Bird
1	128x128	0.1	0	0	0	0.3	89.1	0	0	9.5	1
2	64x64	0	0	0	0	0.2	88.6	0	0	10.2	1
3	32x32	0	0	0	0	0.2	87.9	0	0	10.8	1.1
4	16x16	0	0	0	0	0.2	88.6	0	0	10	1.2
5	8x8	0	0	0	0	0.2	89.7	0	0	8.9	1.2

Table 3.2 Comparison between the results of various image types and sizes (Kekre et al., 2011)

Images	128x128	64x64	32x32	16x16	8x8
Bird	48.8	46.5	44.7	40.3	35.6
Coins	31.7	31.9	30.7	29.5	28.4
Sea	33.9	37.5	37.4	39.9	38.8
Bus	69.2	67.5	65.9	65.1	64.4
Dinosaur	35.3	35.9	37.1	38.4	39.5
Elephant	89.1	88.6	87.9	88.6	89.7
Rose	4.2	5.3	7.9	11.8	18.3
Horse	34.1	33.7	33.4	32.3	30.3
Mountains	46.8	49.6	51.4	51.9	53.6
Rainbow	48.8	49.7	51.5	47.5	41.8

3.3.2 A hybrid DCT, DWT and SVD-based Method for the Extraction of Local Iris Image Features

After preparing the iris image, the method for the selection of the most discriminating features from iris image is apply. The selected features are used for partitioning and classifying the iris templates robustly. The enhanced method is based on two frequency analyses (DCT and DWT) which are simultaneously implemented to obtain the holistic features of the iris image. The SVD was used for reducing the number of features that represent the data.

3.3.2.1 Applying Discrete Cosine Transform (DCT)

This step involves the implementation of DCT transformation on each block. DCT is a common image transformation scheme with several advantages. It is an orthogonal scheme for the transformation of image blocks from space to frequency domains. Additionally, the image is represented by just a few numbers of data points. The generated DCT coefficients can be quantitated with ease; hence, it is possible to achieve a good feature extraction accuracy. The DCT helps in the separation of images into segments of varying importance (Chen *et al.*, 2017). The DCT and the discrete Fourier transform are similar as they transform an image from the spatial to frequency domains, please refer to Figure 3.8.

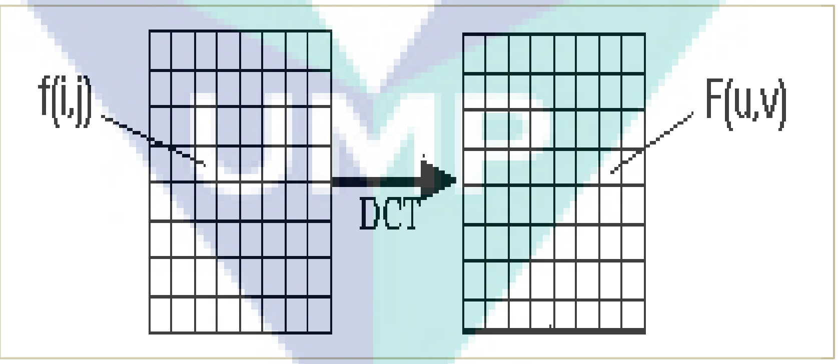


Figure 3.8 Transform image domain using DCT

The DCT uses a set of different block sizes for DCT; however, the transform using each block size takes place independently, resulting in multiple DCT coefficient arrays. To suit these coefficients to the indexing and matching model, they have to be quantized to integers.

3.3.2.2 Applying Discrete Wavelet Transform (DWT)

The next step involves the decomposition of the blocks into different sub-bands (LL 2, HL 2, LH 2, HH 2, HL 1, LH 1 and HH 1) by applying the two level DWT on each of the transformed blocks (Chen, et al., 2017), as shown in Figure 3.9. The advantage of using DWT for block size 8×8 is to reduce the computational difficulties. Multi-resolution decomposition has been observed to provide useful textural discrimination (Sridhar, 2017). The wavelet transform can generally be expressed by Equation 2.13 given in chapter II.

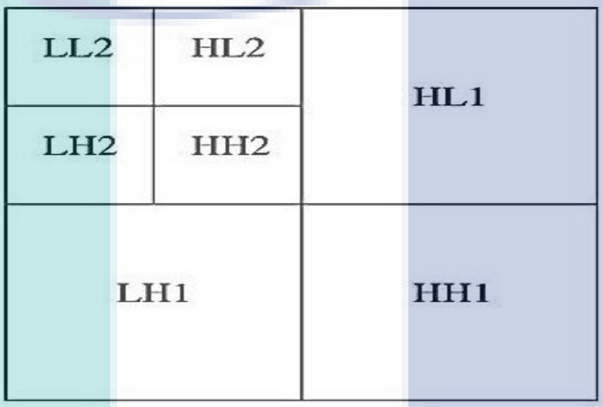


Figure 3.9 2 Level 2D-DWT Decomposition Process

3.3.2.3 Applying singular vector decomposes (SVD)

In this step, the SVD was used, the singular vector decomposes (SVD) applied to a set of different block sizes for DCT for all 2nd level sub band. The two orthogonal unitary matrices are obtained and a diagonal matrix constitutes the singular values of the matrix of sub-bands. In an image, the important feature can be expressed by a series of the singular values along with diagonal direction. In the indexing and matching model, the block size 8×8 image is decomposed into 7 sub bands by 2-level DWT. For uniform saving format, the first two singular values can be applied to the expression features of eac sub band image. So if the researcher has a matrix A with m rows and n columns, with rank r and r ≤ n ≤ m. Then the A can be factorized into three matrices (Nair & Aruna, 2015), Eq. (3.1) and (3.2).

$$A = U\sum V^T \qquad 3.1$$

Where Matrix U = m × m orthogonal matrix, and matrix V = n × n orthogonal matrix

Here, ∑ = m × n diagonal matrix with singular values (SV) on the diagonal with non-negative numbers.

$$A = \underset{mxm}{\begin{bmatrix} | & & | \\ u_1 & \cdots & u_2 \\ | & & | \end{bmatrix}} \underset{mxn}{\begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_p \end{bmatrix}} \underset{nxn}{\begin{bmatrix} - & V_1^T & - \\ & \vdots & \\ - & V_n^T & - \end{bmatrix}} \qquad 3.2$$

Where: $\sigma 1 \geq \sigma 2 \geq \cdots \geq \sigma_p$

3.3.2.4 Singular Vector Feature Selection

In this step the important features $F_{i,j}$ are selected, the 7 sub bands which can be expressed by $\{S1, S2, S3, S4, S5, S6\ and\ S7\}$, the size of $S1$, $S2$, $S3$ and $S4$ is 2×2. The size of $S5$, $S6$ and $S7$ is 4×4. In the existing method, only one value can be applied to energy value for each sub image. Especially, it is inaccurate to express features of energy value for the 4×4 sub band of $S5$, $S6$ and $S7$. In the indexing model, the first two singular values ($\sigma 1$, $\sigma 2$) from ∑ matrix are selected to represents each sub-band see Figure 3.10, where the singular vector expresses the feature for all sub-bands images. Because, the most important feature of an image can be expressed by the biggest singular vector values (Sadygov, 2014).

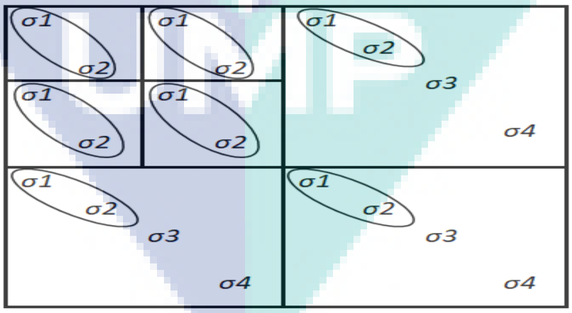


Figure 3.10 Selected singular values of each block

In mathematics particular functional analysis, the singular values, or s-numbers of a compact operator $T : X \rightarrow Y$ acting between Hilbert spaces X and Y, are the square roots of the eigenvalues of the non-negative self-adjoint operator $T^*T : X \rightarrow X$ (where T^* denotes the adjoint of T). The singular values are non-negative real numbers, usually listed in decreasing order $s1(T), s2(T)$ The largest singular value $s_1(T)$ is equal to the operator norm of T (Gohberg & Krein, 1969). The SVD decomposes *X* into three simple transformations: an initial rotation *V**, a scaling Σ along the rotated coordinate axes and a second rotation *U*. *Σ* is a diagonal matrix containing in its diagonal the singular values *σ* of *X*, where the decreasing speed of *σ* is very fast. In the normal case, the sum of the first 10% or 1% of the singular values accounts for more than 99% of the sum of all the singular values. In the proposed method, the former *r* large singular value will be used for describing the approximation of matrix *X*. One data analysis example will be used to present the selection reasons of singular values *σ* in the proposed method as follows. Noise also arises anytime the researcher collect data: no matter how good the instruments are, measurements will always have some error in them. If the researcher remember the theme that large singular values point to important features in a matrix, it seems natural to use a singular value decomposition to study data once it is collected. As an example, suppose that the researcher collect some data as shown in Figure 3.11,

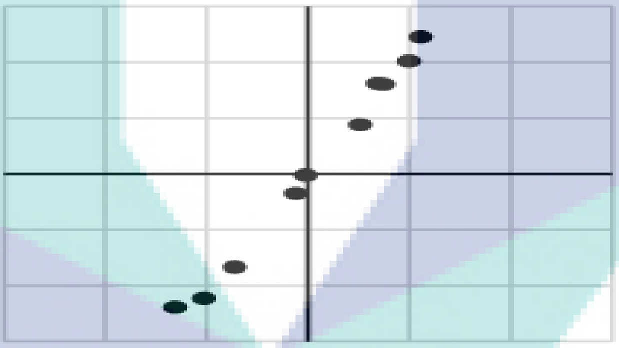


Figure 3.11 Collected data of singular value performance

The researcher may express the data as the following matrix

$$\begin{matrix} -1.03 & 0.74 & -0.02 & 0.51 & -1.31 & 0.99 & 0.69 & -0.12 & -0.72 & 1.11 \\ -2.23 & 1.61 & -0.02 & 0.88 & -2.39 & 2.02 & 1.62 & -0.35 & -1.67 & 2.46 \end{matrix}$$

and perform a singular value decomposition. The researcher find the singular values: $\sigma 1 = 6.04$, $\sigma 2 = 0.22$

With one singular value so much larger than the other, it may be safe to assume that the small value of σ2 is due to noise in the data and that this singular value would ideally be zero. In that case, the matrix would have rank one meaning that all the data lies on the line defined by u_i. please refer to Figure 3.12.

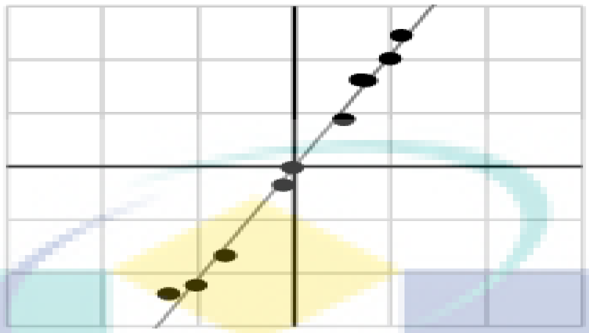


Figure 3.12 All data points lie on the line

This illustration point to the basics of SVD, a set of techniques that uses singular values ($\sigma 1$, $\sigma 2$) to detect data redundancies and dependency. Similarly, SVDs can be deployed for the detection of data groupings. This explains why SVDs are used when attempting to improve feature selection methods. Specifically, the matrix X is decomposed orthogonally so that the vector set of X can be expressed as the projection length σ of each eigenvector U. The next section explains the benefit of combining more than one transformation algorithm for image feature extraction and analysis.

II. The benefit of combining DCT, DWT, and SVD: The reason behind combining the two transformation algorithms DCT and DWT is to take the advantages of both algorithms, where each one has its advantages. DCT technique is a process to modify a signal into elementary frequency components. It is a closely related to discrete Fourier transform (DFT), using the DCT a signal is categorized into its basic frequency components. When the researcher use DCT on X*Y sized matrix, the 2D-DCT extract the energy information of the image and then it will focus on some specific features located in the upper left Corner of the outcome real-valued X*Y DCT matrix Then the result matrix is used as a feature vector technique is used to improve the facial expression images. The recovery of images is practical just because of the DCT. The workings of DCT coefficients return the average energy of pixel blocks whereas the AC components return the intensity of image. As DCT separates an image hooked on discrete blocks of pixels of differing significance or weightage in an image so the researcher can say that DCT is a lossy compression technique.

On the other hand DWT is a mathematical approach that disintegrates a time domain signal into different frequency group. DWT is derived from Continuous Wavelet Transform (CWT) which provides time and frequency information simultaneously (Singh *et al.*, 2016) .The advantages of DWT over other transformation methods is it does not shift and scale endlessly due to its discrete steps nature. Other than that, DWT gives satisfactory details and lessen the computational time significantly. This research will focus on two dimensional DWT. Two dimensional DWT is the process of applying discrete wavelet transform on the rows as well as the columns of an image. Four bands are created when applying two dimensional discrete wavelet transform on the image which are the low-low (LL), low-high (LH), high-low (HL) and high-high (HH) bands (Chen *et al.*, 2017). LL band is created when low pass filter is applied to the horizontal and vertical values of the image. LH band is created when low pass filter is applied to the horizontal and high pass filter is applied to the vertical values of the image. HL band is created when high pass filter is applied to the horizontal values and low pass filter is applied to the vertical values of the image. Lastly, HH band is created when both horizontal and vertical values is filtered with high pass filter (Daqrouq *et al.*, 2017).Mathematically, DWT is defined as shown in Eq. 2.13.

Therefore, combining the DCT and DWT properties have advantages in features extraction, such as the ability to separate information signal into low frequency components uncorrelated with their counterpart frequency index. The DCT coefficients only provide frequency component information without localizing the specific frequencies in space. The DWT solves this problem by analysing the signal at different frequencies and at different times. It should note that the existing methods uses single frequency analysis which is not enough yet to get the holistic information of any iris pose variations. In other word, the DCT still have few missing dominant information of iris image. For combining SVD, the proposed method goal is to reduce the data dimension of analysed image in order to selecting features, so can use for classification and indexing. The data can be described with fewer dimensions, without losing much of the meaning of the data. The data reside in a space of lower dimensionality essentially, the researchers assume that some of the data is noise, and the researcher can approximate the useful part with a lower dimensionality space. Dimensionality reduction does not just reduce the amount of data; it often brings out the useful part of the data. Singular value decomposition (SVD) is a technique that allows an exact representation of any matrix (Zgrzywa *et al.*, 2017). SVD can represent a high-dimensional matrix into a low-dimensional by eliminating the less important parts and produce an approximated representation with any desired number of dimensions (rank).

3.3.3 Creating feature vector of iris image

After extracting the features of all image blocks, each block has two features from each subband, The singular values are grouped together by summing up as (Mehrotra *et al.*, 2009):

$$F = \sum_{i=1}^{b} S_i(x, y) \qquad 3.3$$

where t is the number of blocks, and S is the singular value $\sigma 1$ for the image (x,y). The same process for the singular value $\sigma 2$, as shown in Figure 3.13. The result of this process is a seven of two-dimensional features $F_{i,j}$; each pair of feature represents a key used for the partitioning process.

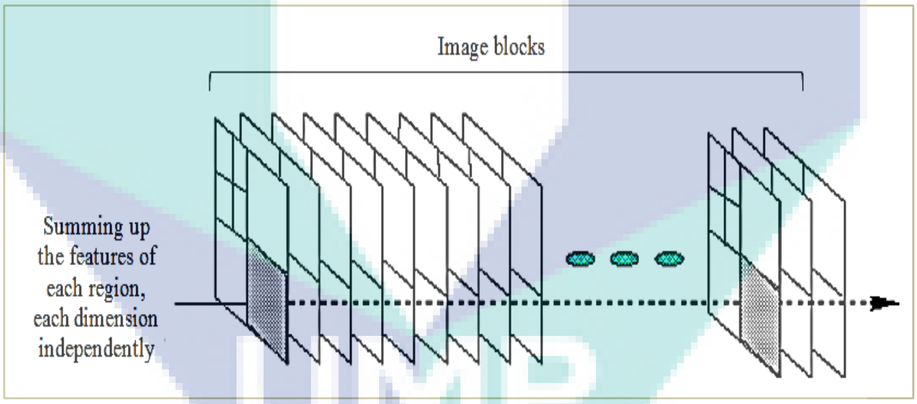


Image blocks
Summing up the features of each region, each dimension independently

Figure 3.13 Feature vector creation for iris image

The flowchart of the extraction method is depicted in Figure 3.14. The steps of the enhanced approach are as follows:

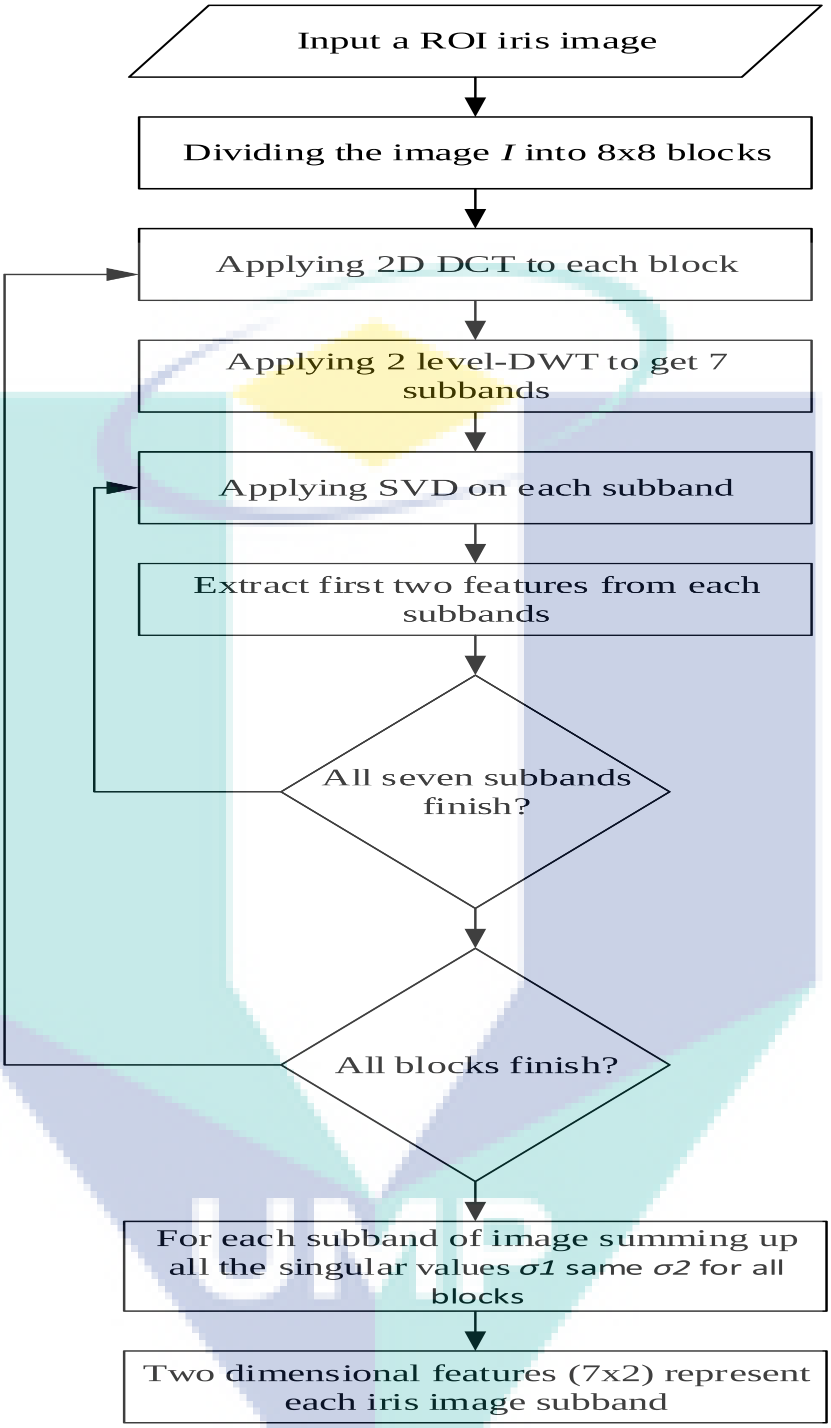

Input a ROI iris image
Dividing the image *I* into 8x8 blocks
Applying 2D DCT to each block
Applying 2 level-DWT to get 7 subbands
Applying SVD on each subband
Extract first two features from each subbands
All seven subbands finish?
All blocks finish?
For each subband of image summing up all the singular values σ1 same σ2 for all blocks
Two dimensional features (7x2) represent each iris image subband

Figure 3.14 Extraction of local features from an iris image

3.4 Clustering Phase

3.4.1 Partitioning based on Scalable K-means++ Algorithm

Database partitioning into sub-bands is the next process. This partitioning is done to reduce the solution space. Clustering refers to object partitioning into different clusters. The query response time is expected to depend on the similarity of the query template to the database template and not on the total number of templates contained in the database. Hence, there is a need to partition the database using a clustering method such that images with similar textures can be clustered (Mehrotra, 2010). In the enhanced method, the features of each sub-band of all the images are grouped together based on similarity to get seven kinds of groups represented by seven sub-bands, shown in Figure 3.15.

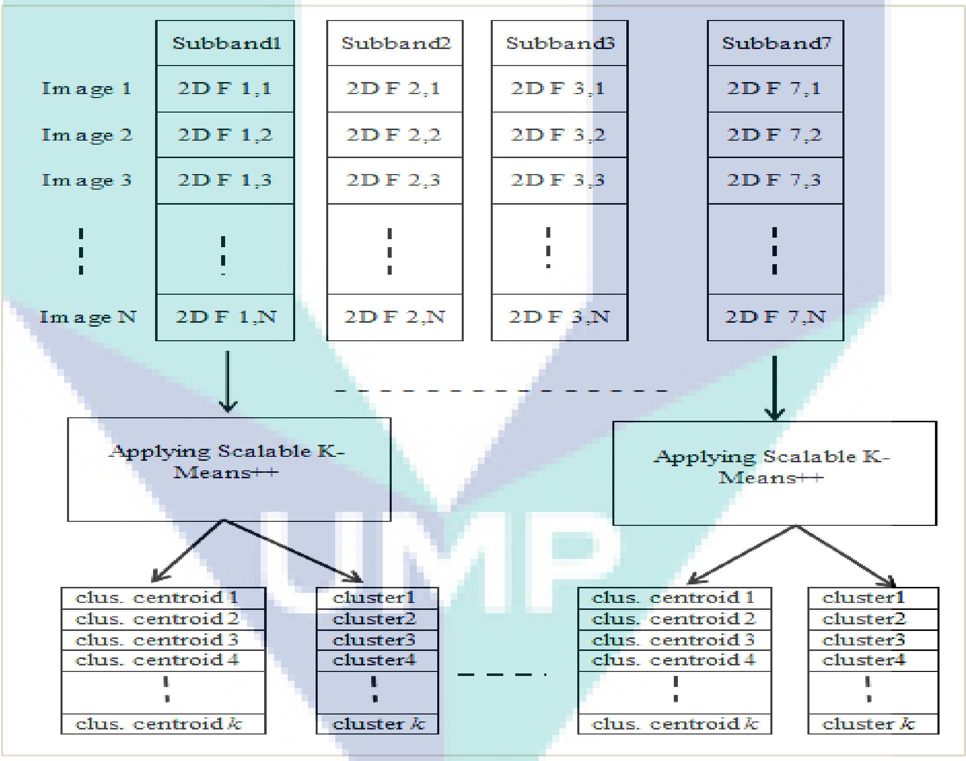


Figure 3.15 The partitioning and clustering approach

The scalable k-means++ (K-means||) is used in the improved approach to partitioning the local features into groups. This algorithm iteratively assigns data points to *K* groups based on the provided features, while data points clustering are based on feature similarity. The outcomes of K-means clustering are:

i. The cluster *centroids* can serve as labels for new datasets.

ii. Each data point is clustered.

The resulting *centroids* are used for clustering when there are new features, effectively transforming the initial set of unknown data points into dataset points with class identities. Instead of establishing the groups before considering the data, clustering permits the analysis of the originally formed groups. The following sections explained the benefit of using scalable K-means++ instead of the traditional K-means.

3.4.1.1 The benefit of using Scalable K-means++ (K-means||)

K-means is a common clustering algorithm with several advantages such as speed and ease of use. These advantages make it possible to run the algorithm on large datasets (Raykov *et al.*, 2016). A major problem of the K-means is its inability to give the same results in each run. This is due to the dependence of the resulting clusters on the initial random assignments. Besides, it is highly sensitive to initial seed selection of cluster centers (Burks *et al.*, 2015). In the K-means++, the initial k-means centres are proposed to be seeded by creating a simple probabilistic approach for the generation of the initial k-means clustering centres. The basic idea of k-means++ is that the initial centroid should be far away from each other. The algorithm starts by randomly choosing a centroid c_0 from all data points. For centroid c_i, the probability of a data point x been chosen as a centroid is proportional to the squares of the distance of x to its nearest centroid. In this way, k-means++ always tries to select centroids that are far away from the existing centroids, which leads to significant improvement over k-means with a bit sacrifice on the runtime.

Unfortunately, K-means++ is inefficient while handling large datasets. As the dataset grows, the number of classes increases. This condition further complicates the distance calculation procedure.The K-means|| was proposed to address the issues of scalability in K-means++. It was developed as a parallel version of K-means++. A parallel version of K-means++ method is implemented in Apache Spark. The method is based on the idea of sampling $O(k)$ points in each iteration rather than one single iteration. This approach achieved a better performance compared to K-means++ both in parallel and sequential settings. It has a faster speed compared to K-means and GMM as it exploits the computational advantages of the MapReduce model (Bahmani *et al.*, 2012). K-means|| is a modified version of the K-means++ in which noisy outliers are managed to speed up the computation by reducing the number of iterations.

3.4.1.2 Distance measurement

The distance between two series can be used to determine their similarities. A number of such distances exist for the measuring of the similarities between series. Among such distance metrics, the most adopted is the Euclidean Distance (ED). The next section explains the analysis of the results of implementing the k-means clustering with various distance functions such as the Euclidean Distance function (ED) and Manhattan distance function (MD).

I. The K-means clustering algorithm with various distance functions: This study aims to analyze the results of implementing the k-means clustering with various distance functions for performance evaluation. The k-means algorithm was used in this study together with The Euclidean distance and Manhattan distance function. The outcome of these implementations was analyzed and compared (Figure 3.16). The experiments were performed on a dummy dataset with the following description:

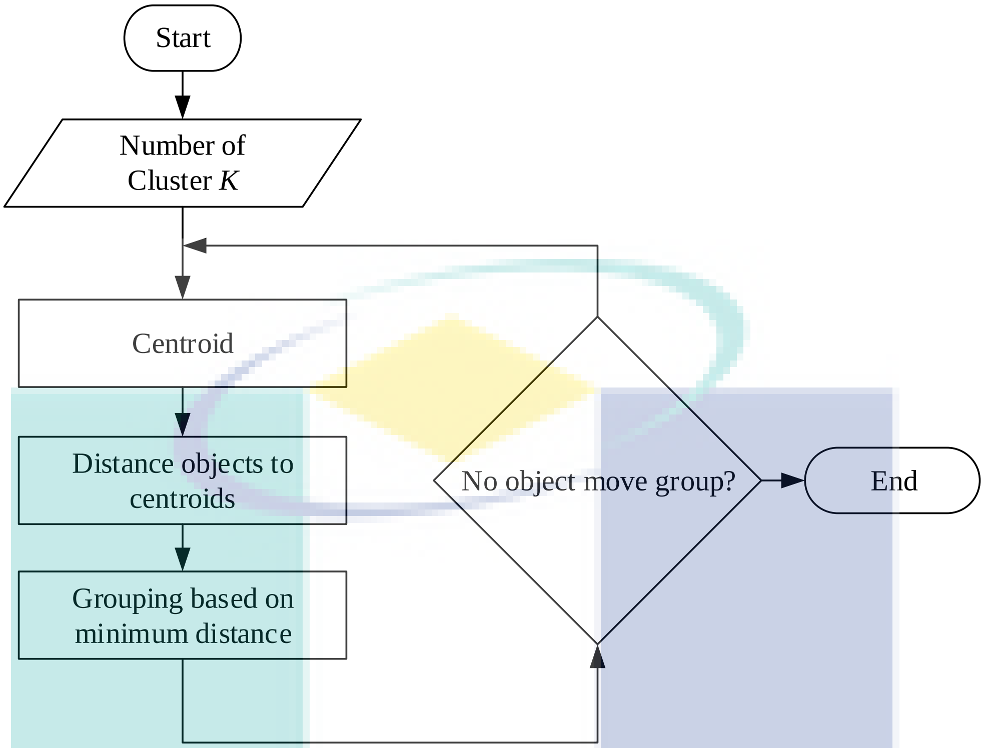

Start
Number of Cluster *K*
Centroid
Distance objects to centroids
Grouping based on minimum distance
No object move group?
End

Figure 3.16 Steps of K-means clustering

The dataset used consists of 200 instances and their respective activities. The data is composed of several attributes that represented their online behaviour. WEKA was used to implement this clustering process. WEKA supports numerous data mining operations such as data pre-processing, clustering, visualization, clustering, feature selection, and regression (Singh & Yadav, 2013). The comparison of the performances of the k-means algorithm using either Euclidean distance (ED) or Manhattan distance function (MD) with various parameters such as the number of iterations is shown in the following figures. Also shown are the sum squared errors and required time to build each model. Figure 3.17 shows that when using k-means with Manhattan distance function, the number of clusters tends to increase with the number of iterations compared to the use of Euclidean distance function.

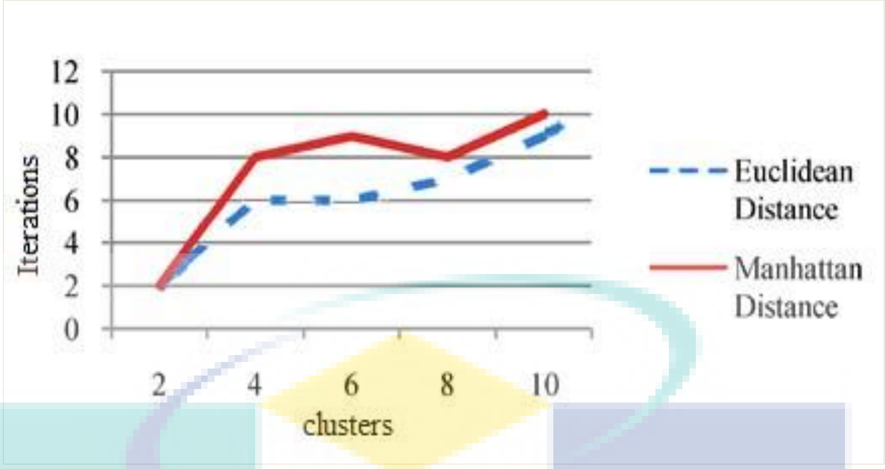


Figure 3.17 No. of clusters vs. no. of iterations

Figure 3.18 depicts the required time to build the model when using k-means clustering with Euclidean distance and Manhattan distance function. The figure showed that it takes fewer time to build the model using K-means with Euclidean distance compared to using k-means with Manhattan distance function. From the result analysis, it was found that Euclidean distance outperformed the other distance functions when used with K-means clustering.

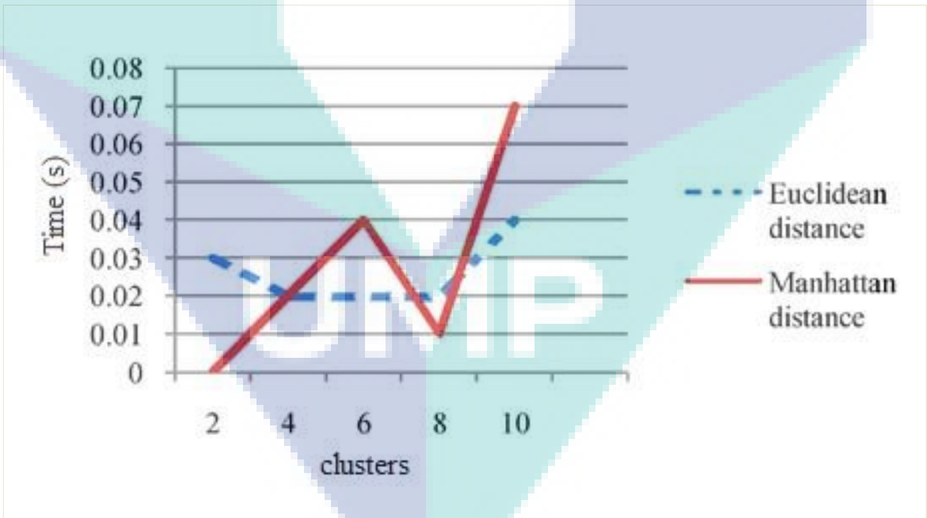


Figure 3.18 Number of clusters against the required time to build a model

3.4.2 The Improved Clustering Method (WKIFA) for Accurate Clustering and Indexing

Clustering is an important part in the whole research steps, shouldering on the responsibility of determining the partitioning and classification success. Clustering is an unsupervised learning technique used for classification of data. Data elements are partitioned into groups called clusters that represent proximate collections of data elements based on a distance function. The aim is to find an optimal partitioning for a group of unknown data. Here, an improved clustering algorithm is presented which combines the K-means and Fireflies Algorithm (FA). The K-means algorithm is sensitive to initial cluster and data noise, etc. To overcome these shortages, it has been used the Fireflies Algorithm (FA) which has power ability of global search and quick convergence rate to optimize the initial clustering centers of traditional K-means algorithm. As the same time, a kind of weighted Euclidean distance also presented to reduce the defects made by noise data and other uncertainties, next sections explain the details of the improved method.

3.4.2.1 Weighted K-means Algorithm Based on Improved Firefly

Traditional k-means clustering partitions a group of objects into a number of non-overlapping sets, which is based on distance partitioning clustering algorithm. It uses Euclidean distance (ED) division criteria which has some limitations that can lead to weak clustering. The main drawback of k-means is that the initial clustering center sensitivity, noise and the existence of abnormal points will have a certain impact on the quality of the cluster. The basic idea of this new method is to obtain the optimal solution as the K-means algorithm based on the initial clustering center. It considered the impact of sample data differences, the introduction of traditional Euclidean distance (ED) with weights, increasing the distinction between data attribute levels, and reducing the impact of abnormal points.

The benefit of weight is to make features more distinction and the mathematical process more concentrated and complex, if the researcher has one feature have same distance to two centroids, the added weight will distinction between the two distances, for better clustering. The perturbation method has been improved to achieve global optimization and improve clustering accuracy and results in stability.

3.4.2.2 Weight Calculation

The goal of this weight is to make features are more distinction, and make the mathematical process more concentrated and complexity if the researcher have one feature have same distance to two centroids, the added weight will distinction between the two distances, for better clustering. The equation 3.4 is to calculate the proposed weight, if there are n sample data, $X = \{x_1, x_2, x_3 \ldots x_n\} \epsilon R^m$ is the image features data to be clustered, while $x_i = (x_{i,1}, x_{i,2}, x_{i,3} \ldots x_{i,m)})^T$ are m-dimensional vectors. In the data, the influence of each component is different and the weight is defined as:

$\omega = [\omega_1, \omega_2, \omega_3 \ldots. \omega_n]^T \epsilon R^{n\times m}$, where: $\omega = (\omega_{i1}, \omega_{i2}, \omega_{i3} \ldots. \omega_{im})^T$ Is an m-dimensional vector. The weights are defined as:

$$\omega_{id} = x_{id} / \frac{1}{n} \sum_{d=1}^{n} x_{id} \tag*{3.4}$$

where x_{id} represents the first component in the first sample, $\frac{1}{n}\sum_{d=1}^{n} x_{id}$ indicates sample. The average of the sum of the d th components of each data object in the data, ω, shows the overall distribution of the sample data.

3.4.2.3 Weighted objective functions

The traditional K-means is used in finding a set of clustering centers $V=\{v_1, v_2, v_3 \ldots v_k\}$ which minimizes the objective function as:

$$J(X,V) = \sum_{j=1}^{k} \sum_{x_i \in G_j} d(x_i, v_j) \tag*{3.5}$$

where: G_j on behalf of the first j is a collection of samples in a category v_j right G_j within all samples x_i of the cluster center, $d(x_i, v_j)$ indicates sample data x_i, and clustering center v_j between the Euclidean distance (ED), defined as follows:

$$d(x_i, v_j) = \|x_i - v_j\| = \sqrt{\sum_{l=1}^{m} |x_{il} - v_{jl}|^2} \tag*{3.6}$$

$$v_j = \frac{1}{n_j} \sum x_i \tag*{3.7}$$

where: v_j denotes j the center of the class, $i = 1, 2, 3 \ldots n$, $j = 1, 2, 3 \ldots k$, n_j is a class v_j in the number of sample data x_i on behalf of v_j in the sample data. For the traditional Euclidean calculation method, the weight is introduced ω; the relationship between each sample data and cluster centers is:

$$d_\omega(x_i, v_j) = \|x_i - v_j\|_\omega = \sqrt{\sum_{j=1}^{m} \omega_{id} |x_i - v_j|^2} \qquad 3.8$$

The objective function is shown in Equation 3.9.

$$J_c(X, V) = \sum_{j=1}^{k} \sum_{x_i \in G_j} \sqrt{\sum_{j=1}^{m} \omega_{id} |x_i - v_j|^2} \qquad 3.9$$

After introducing the attribute weights in Equation 3.6, the traditional K-means algorithm is not changed, and the distance between the normal data and the clustering center is smaller, The distance between the center of the class and abnormal data becomes larger, and makes the original distribution is not obvious, and the data samples are not easy to be classified improved objective function become more prominent and more suitable for the clustering. After the improvement not only improve the clustering accuracy, but also can effectively reduce the number of iterations algorithm to improve efficiency.

3.4.2.4 Improved firefly algorithm calculation of attractiveness and disturbing way

The improved degree of attraction is calculated as:

$$\beta(r) = \frac{\beta_0}{1+\gamma\times r_{ij}^2} \qquad 3.10$$

When the firefly j attracts the firefly i to itself, the distance between them will be reduced. Using the equivalent infinitesimal replacement principle in higher mathematics, the simple fraction of the right side of the equal sign is used instead of the exponential function on the left side. The computational complexity is smaller and easier to implement, $e^{-\gamma x r_{ij}} \approx \frac{1}{1+\gamma\times r_{ij}^2}$. The attractiveness of the firefly is calculated by equation 3.10, and the Equation for updating the firefly position is shown in Eq. 3.11 (Tilahun & Ong, 2012):

$$x_{i+1} = x_i + \frac{\beta_0}{1+\gamma\times r_{ij}^2} \times (x_i - v_0) + a \times (rand - 0.5) \qquad 3.11$$

where: v_0 is the current optimal clustering center, β, γ, α, *rand*, with meanings, as earlier described.

$$v_0 = \frac{1}{n_i}\sum_{y\epsilon\Gamma_i} y \qquad 3.12$$

where n_i is the first Γ_i number of data in the cluster, *y* represents the data value.

In the traditional firefly algorithm, the fireflies are attracted to the relatively bright firefly movement, where the disturbance items α×(rand-0.5) increase the algorithm search area, although to avoid premature fall into the local optimal, increase the algorithm's local search ability, but eventually lead to the whole strategy convergence speed is slower, the stability is relatively poor. When the objective function value to be processed is relatively large, the perturbation effect is not obvious, and the algorithm is easy to cause fluctuation near the local optimal value.

I. Location of fireflies update

Several experiments have been performed to obtain the following formula in order to generate a new solution by change the constant value (*rand*-0.5) into variable value ($rand\ x\ (xi\ -\ v0)^2$) to make the random movement depends on the resulted distance. After introducing the perturbation operator $\alpha \times rand \times (x_i - v_0)^2$, the position update equation of the firefly can be expressed as:

$$x_{i+1} = x_i + \frac{\beta_0}{1+\gamma\times r_{ij}^2} \times (x_i - v_0) + a \times rand \times (x_i - v_0)^2 \qquad 3.13$$

Eq. 3.13 is based on the optimal class center which is also a random disturbance. The brightest firefly moves according to Eq. 3.14.

$$x_{*+1} = x_* + a \times rand \times (x_i - v_0)^2 \qquad 3.14$$

Where x_* represents the location of the brightest fireflies currently, v_0 same as above. In the firefly algorithm, according to Eq. 3.13, the random movement of the brightest firefly can be avoided and this can easily lead to the stagnation of the algorithm in the local optimal value, or the global and convergence speed will be slow.

II. Basic idea of weighted K-means algorithm based on improved firefly (WKIFA)

The basic idea of the firefly algorithm: using fireflies to represent the solution to the clustering problem, where it represents the quality of the solution. Between fireflies through constant information exchanging, they attract each other, move, gradually gathered, the final "group" points, to find the best clustering situation condition. Where the location of the fluorescent brightness is large, which represents the cluster center, which the firefly have the greater brightness and the stronger the attraction and it indicates the better the location of the firefly. Through the firefly the mutual attraction and movement between the firefly come to find the optimal clustering center with the goal of making it that the sum of the distances from the sample to the center of the corresponding cluster is minimal. Firefly Brightness I with objective function $J_{c.}$

$$I = J_c \qquad 3.15$$

The brightness value of the firefly is closely related to the size of the objective function value The size reflects the location of the firefly is good or bad, the smaller the objective function value. The higher fireflies brightness. In this thesis, the objective function value is used to calculate the brightness, The smaller the time, and the time complexity of the algorithm is reduced, clustering the smaller the value of function Jc, the better the effect of the entire cluster.

III. Basic steps of weighted K-means algorithm based on improved firefly (WKIFA)

Step 1: Initialization parameters: Given the cluster number k, population size N, maximum attractiveness, light intensity $\beta_{0,}$ absorption coefficient γ, step factor α, maximum iterations number $T_{\max}$, iteration stopping threshold ε;

Step 2: Randomly select a point k as the initial position of the firefly, then, determine the distance from each point to the cluster center using Eq. 3.8;

Step 3: According to the calculated distance, the sample points are sequentially divided into the category where the cluster center is closest to it;

Step 4: According to the initial clustering results obtained in Step 3, the brightness of the firefly is calculated according to Equation 3.9.

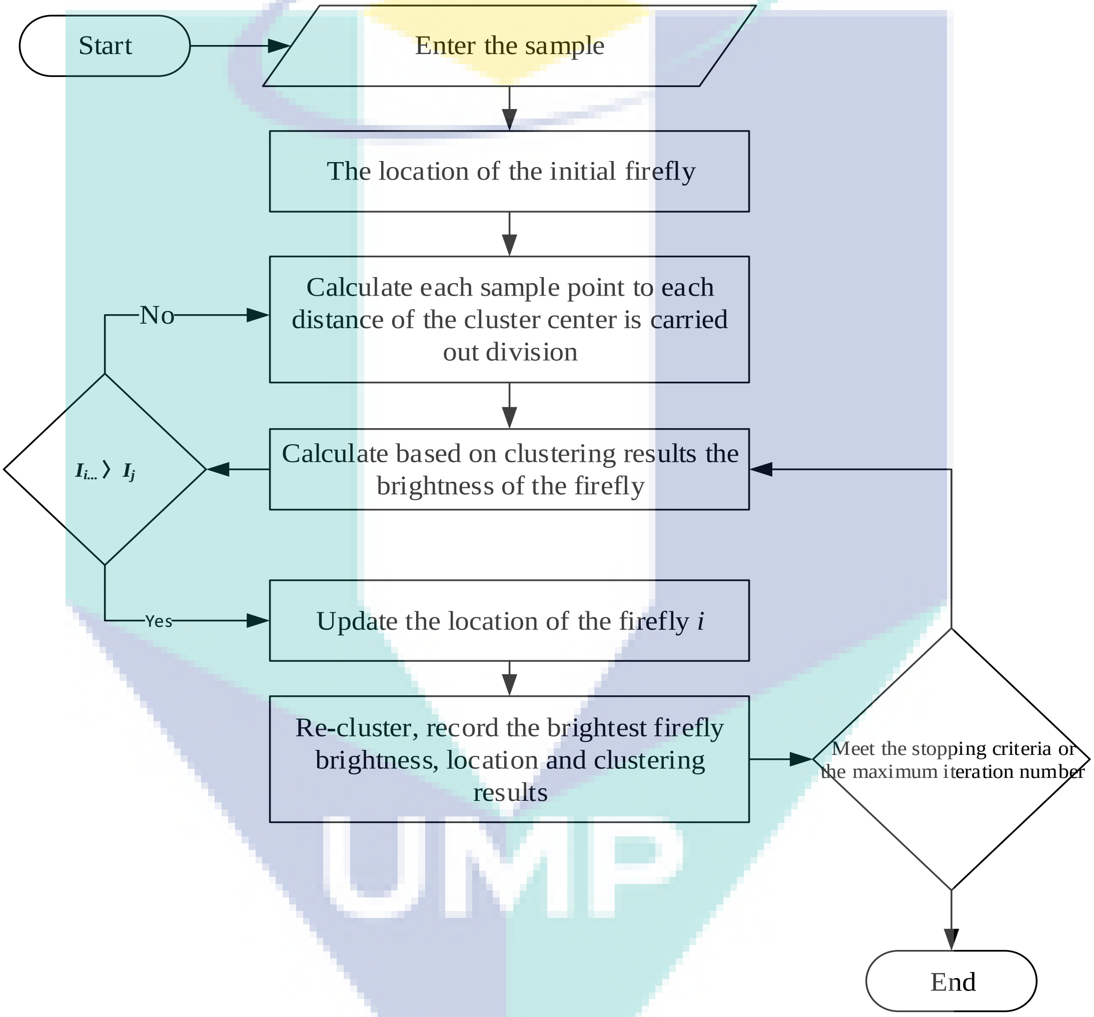
Step 5: If $I_i>I_{j,}$ the firefly's objective function j is small. This means that the firefly is in a good location and the firefly j will attract the fireflies i to move it. The size of the movement is determined by the Equation 3.10, 3.13 and 3.14 to update the location of the firefly;

Step 6: Update the location, re-clustering, and update the firefly brightness, re-record the brightest firefly brightness, location, clustering results;

Step 7: Reaches the maximum number of iterations or reaches the threshold, the algorithm is stopped, otherwise go to step 4;

Step 8: Output the results.

The basic algorithmic flowchart is presented in Figure 3.19.



Start
Enter the sample
The location of the initial firefly
Calculate each sample point to each distance of the cluster center is carried out division
No
Calculate based on clustering results the brightness of the firefly
$I_{i...} \rangle \ I_j$
Yes
Update the location of the firefly *i*
Re-cluster, record the brightest firefly brightness, location and clustering results
Meet the stopping criteria or the maximum iteration number
End

Figure 3.19 The clustering method (WKIFA)

After the algorithm is improved, the sample points converge to the peak of the function, and the convergence effect is better. The results in ch4 shown that the clustering center point of this algorithm can get more accurate and stable clustering.

3.5 Search Approach Phase

3.5.1 The Searching Approach Based on parallelization, B-tree and Half-searching

A common problem that usually emerges in the large biometric database is the complexity of searching the whole database when trying to retrieve an identity. In most cases, the input data is matched against all the data contained in the database. Another problem is the increased rate of FAR as the size of the database increases (Dey & Samanta, 2012). Consequently, this affects the search response time, accuracy, and retrieval efficiency of the system. According to the existing methods, the indexed groups need to search for query image linearly. The performance and efficiency are acceptable with the small database size. However, for the large database size, the searching time seems to be not efficient. Especially, in the hardware designs and practical application, linear searching is not the optimal solution.

An efficient searching approach is proposed which complement our partitioning and clustering method for better retrieval efficiency. The searching approach is consisting of indexing the database, database arrangement and searching method

3.5.1.1 Indexing the Database

The index of a database refers to the data structure which can improve the speed of data retrieval processes. Indexes are deployed for a quick data location without the need of searching all the data in the database (Pan *et al.*, 2014). In this way, the indexing of biometrics database is done using 2D features vector of each image stored in the database. The obtained feature vector of each image contains 7 pairs of values from each subband. Each subbands' feature is partitioned into logical groups where iris strips with similar texture details are placed in the same group for a better and more accurate matching.

I. Global key generation

The key is used for inserting an image in the database at the time of enrolment. The data structure using to store an iris template is B tree, where the generated global key of each image is used to traverse the b-trees. The images with same key are stored in the same nodes. For a given query image, the key is generated and tree is traversed to end up to a node. The templates stored at the node are retrieved and compared with the query template to find the best match. Each key consists of seven pair of values which represents the images' subbands. Each key number represents the group number that the feature pair belongs to. For example the key as (2-5-1-4-2-3-9) the first value is group 2 this group which is include the first pair of subband features in while 5 is include the second pair and so on. All the key numbers in each subband are combined following the Morton Order Traversal (Mehrotra *et al.*, 2009). In this way, the low-frequency coefficients are placed before those with high frequency. This is done with all the images in the database to obtain the keys. Figure 3.20 is a schematic representation of the Morton Order Traversal.

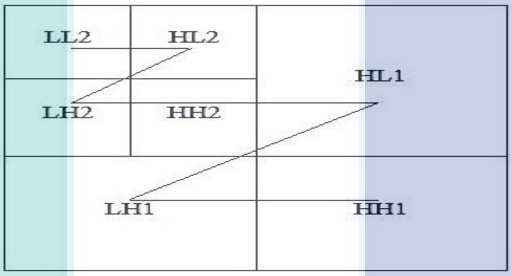


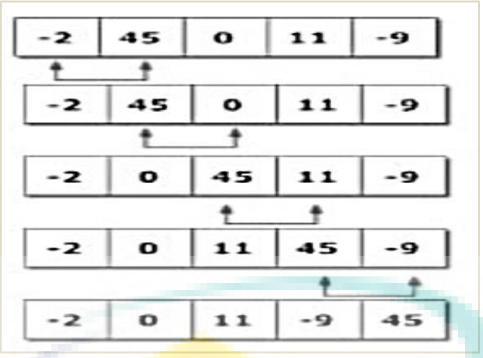
Figure 3.20 Morton order-based traversal

II. Database creation

The database creation consists of the following steps:

i. Features ranking algorithm

Based on the improved partitioning method, all the feature data are divided into groups. To prepare the database for searching processes, all the elements of each group are ranked in an increasing order. This ranking is important for optimizing the searching algorithm. It is done based on the first singular value σ_1. The ranking process is carried out through the comparison of the first 2 elements of an array. These elements may be swapped if necessary, such as when the first element is greater than the second one. Similarly, the first and third elements can be swapped (if necessary) after comparison. This process continues until the last element has been compared with the first element (Figure 3.21). This step is known as the sorting stage. To achieve a better performance, the comparison of the elements should be initiated from the second element since after the first step, the required number is placed at the first automatically.



-2	45	0	11	-9
-2	45	0	11	-9
-2	0	45	11	-9
-2	0	11	45	-9
-2	0	11	-9	45

Figure 3.21 Example of an ascending ranking process

ii. Distributing the database

Searching images in a large database may still take a long time even with efficient indexing and partitioning approaches. To address this computational challenge, a parallelization search strategy has been used. The stored groups which include the pairs of features (*σ1*, *σ2*) are divided into two bins (bin1 includes *σ*1 features and bin2 includes *σ*2 features) as shown in Figure 3.22. Each of the Bin1s and Bin2s forms two B-tree structures as shown in Figure 3.23.

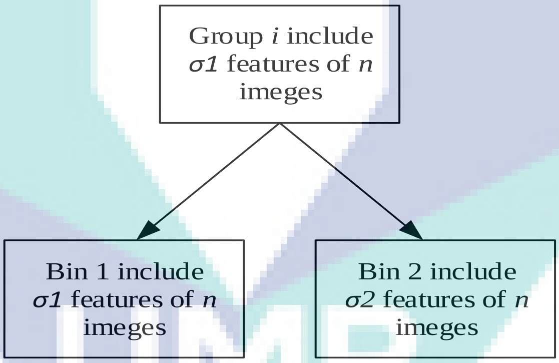
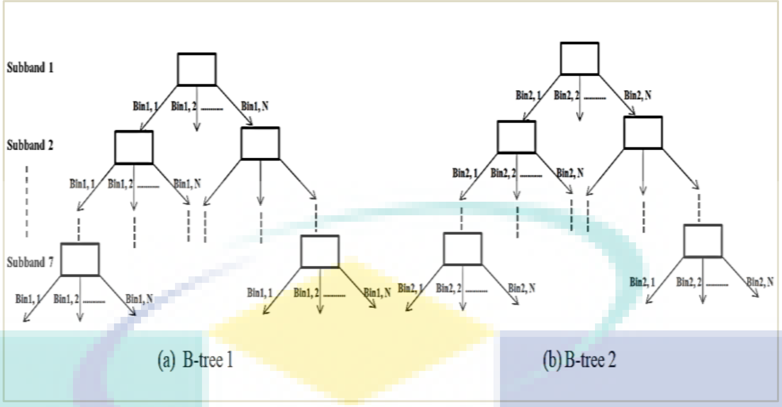

Group *i* include *σ1* features of *n* imeges
Bin 1 include *σ1* features of *n* imeges
Bin 2 include *σ2* features of *n* imeges

Figure 3.22 Dividing each group into two Bins



(a) B-tree 1

(b) B-tree 2

Figure 3.23 The two B-tree structures

To get two B-trees, the overall number of bins constructed for each subband represents the degree of the tree. The tree height is represented by the number of subbands considered. The trees' root node represents the subband S_0 while the bins are the children. The branch on the left side represents the first bin while the second bin represents the next branch, etc. Each node contained in the second tree level represents the associated subband. The B-tree is traversed with the generated global key, where both trees have the same arrangement. The location of σ1 in the first tree is the same location of *σ2* in the second tree; so, both trees use one key at the same time for traversing processes.

After ranking all the elements of the groups, the basic steps of database distributing as follow:

Step1: Divide each group into 2 bins with feature components as earlier described.

Step3: Distribute the database into 2 B-trees

Step4: Use the global key to simultaneously traverse the 2 B-trees.

Step5: Store the image at the leaf node of the 2 B-trees.

3.5.1.2 Searching Method

The optimal match to search images is achieved by a holistic database search using a global key. The features are extracted firstly using the steps highlighted in section 3.4 before finding the closest centroid for the feature and classifying them into bins. The key represents the bin number it belongs, and it is used to traverse the two trees in parallel. A complete or partial key can be used to traverse the two trees (Mehrotra, Srinivas, Majhi & Gupta, 2009). The searching approach is shown in Figure 3.24.

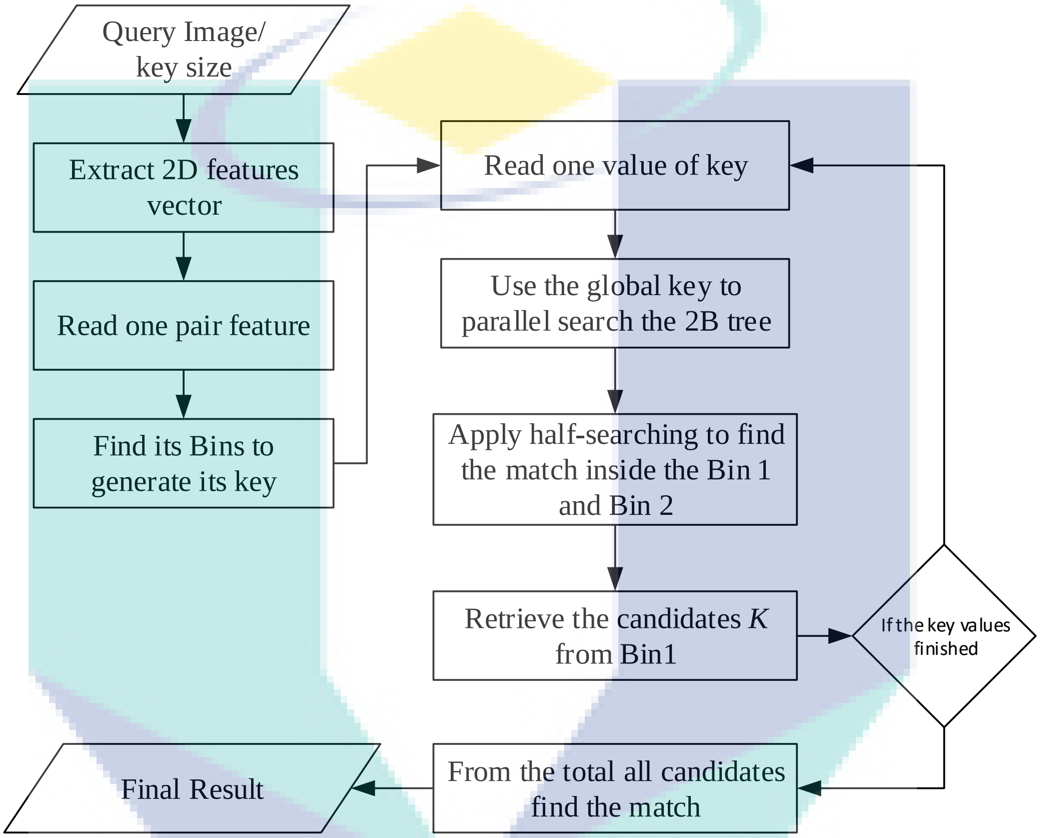

Query Image/ key size
Extract 2D features vector
Read one pair feature
Find its Bins to generate its key
Read one value of key
Use the global key to parallel search the 2B tree
Apply half-searching to find the match inside the Bin 1 and Bin 2
Retrieve the candidates K from Bin1
If the key values finished
From the total all candidates find the match
Final Result

Figure 3.24 The searching approach

I. *Half-searching Algorithm*

The global key is used to reach the bin location (Bin1 in tree1 & Bin2 in tree2) in the existing method when searching linearly (sequentially). This method works well with small bin sizes, but when the deal is with a large data bin size, the traditional searching method becomes inefficient. The half-searching algorithm is used to find the specific feature location inside the bin in order to keep the arrangement of the data in each bin where the features are ranked in an ascending order (section 3.6.1.1 step1). This step is not included in the existing methods as the search is done linearly. With half searching, the searching time for elements in a sequence is reduced by half. This reduction is implemented based on two comparisons;

i. To decide whether to search the upper or lower half of the sequence.

ii. To determine if the sequence has elements.

If *f*(*n*) is the required number of comparisons to search an element in a sequence of size *n*, then,

$$f(n) = f(n/2) + 2 \qquad 3.16$$

when *n* is even.

II. Description of the algorithm

Half searching algorithm is shown in Figure 3.25. The steps are as follows:

Firstly, determine the middle of the entire search interval as follow:- Secondly, compare the value of the keyword to be checked with the keyword value in the middle position. If they are the same, then find the success. However, if they are more, then continue the half-search in the back half of the area. On the other hand, if they are lesser, then, continue the half-search in the former half of the area. Thirdly, repeat the above steps for the reduced area and, then, apply the half-searching Eq. 3.17. Finally, get the result (either successfully or unsuccessfully found).

$$mid = (left + right) / 2 \qquad 3.17$$

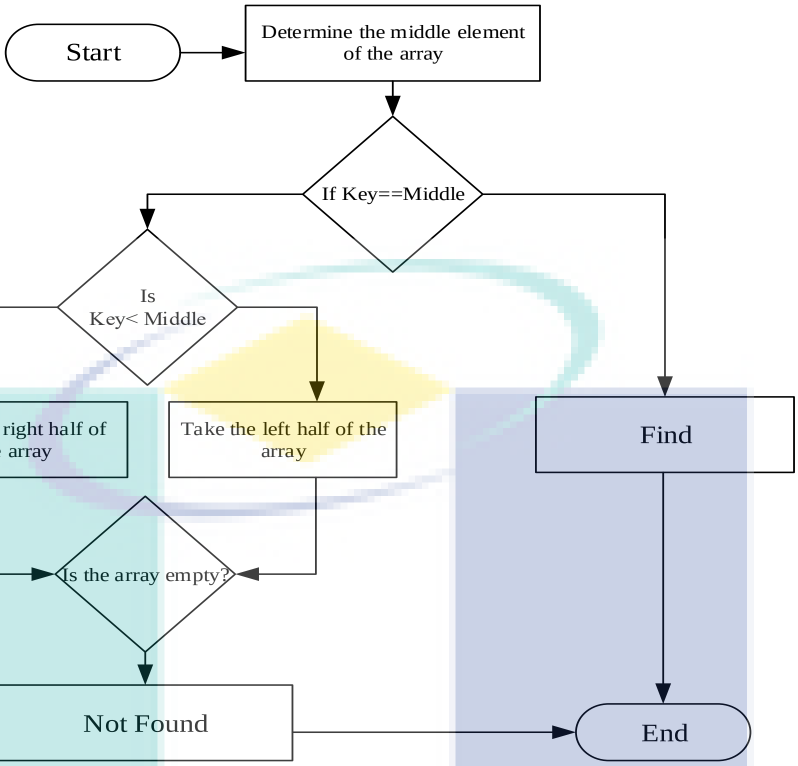

Start
Determine the middle element of the array
If Key==Middle
Is Key< Middle
Take the right half of the array
Take the left half of the array
Find
Is the array empty?
Not Found
End

Figure 3.25 Half-searching algorithm

III. Analysis of half-searching algorithm

To calculate the average case running time of half-searching on n sorted elements, the researcher first assumes that:

- The value being searched for is in the array.
- Each value is equally likely to be in the array.
- The array size is n = 2k-1, with k = a positive value.

Firstly, the researcher observes that using one comparison:- the researcher can find 1 element. If the researcher uses two comparisons, there are two possible elements the researcher can find. In general, after using k comparisons, the researcher can find 2k-1 elements. (To see this, consider doing a binary search on the array 2, 5, 6, 8, 12, 17, 19. 8 would be found in 1 comparison, 5 and 17 in two, and 1, 6, 12 and 19 would be found in 3 comparisons). The expected number of comparisons the researcher makes when running the algorithm would be a sum of the number of comparisons necessary to find each individual element multiplied by the probability of the researcher searching for that element. Let *p(j)* represent the number of comparisons it would take to find element j, then, the sum the researcher have is:

$$\sum_{j=1}^{n} \frac{1}{n} p(j) = \frac{1}{n} \sum_{j=1}^{n} p(j) \qquad 3.18$$

Now, the trick will be to determine that sum. However, the researcher have already outlined that *p*(*j*) will be 1 for one value of j, 2 for 2 values of *j*, 3 for 4 values of *j*, etc. Since *n*=2*k*-1. Hence, the researcher can formulate the sum as follows:

$$\sum_{j=1}^{n} \frac{1}{n} p(j) = \frac{1}{n} \sum_{j=1}^{n} p(j) = \frac{1}{n} \sum_{j=1}^{k} j2^{j-1} \qquad 3.19$$

This is because the value *j* appears exactly 2^{j-1} times in the original sum.

The determined sum can then be expressed as follow:

$$\sum_{j=1}^{k} j2^{j-1} = 1(2^0) + 2(2^1) + \cdots + k(2^{k-1}) \qquad 3.20$$

$$-2\sum_{j=1}^{k} j2^{j-1} = 1(2^1) + 2(2^2) + \cdots + (k-1)(2^{k-1}) + k(2^k) \qquad 3.21$$

Subtracting the bottom equation from the top, the researcher gets the following:

$$-\sum_{j=1}^{k} j2^{j-1} = 2^0 + 2^1 + 2^2 + \cdots + 2^{k-1} - k2^k \qquad 3.22$$

$$-\sum_{j=1}^{k} j2^{j-1} = 2^k - 1 - k2^{k2} \qquad 3.23$$

$$\sum_{j=1}^{k} j2^{j-1} = -2^k + 1 + k2^k \qquad 3.24$$

$$\sum_{j=1}^{k} j2^{j-1} = (k-1)2^k + 1 \qquad 3.25$$

Thus, the average run-time of the binary search is

$$\frac{(k-1)2^k+1}{n} = \frac{(K-1)2^k+1}{2^k-1} \approx k-1 = O(logN) \qquad 3.26$$

In computer science, the three cases of a given algorithm the best, worst, and average represent what the resource usage is at least, at most and on average, respectively. Usually the resource being considered is running time, i.e. time complexity, but it could also be the memory or other resource. The best case is the function which performs the minimum number of steps on input data of *n* elements. On the other hand, the worst case is the function which performs the maximum number of steps on input data of size n whereas the average case is the function which performs an average number of steps on input data of *n* elements. From the complexity analysis, it was observed that the average case run-time of the half search algorithm is much closer to the worst case run-time than the best-case runtime. As such it was found that the worst case number of comparisons is *k* with the average number of comparisons k-1, where *k* = *O*(log n)). The best-case running time of a half searching of *n* elements is *O*(1), corresponding to what was observed for the element of interest in the first comparison. It has been reported that the complexity of the sequential algorithm which is used in the existing methods is *O*(n) in the worst case and the average case, both for successful and unsuccessful search (Karthick, 2016). Thus, half search is more efficient than sequential search; it has a time complexity of *O*(log n) in the average-case running time.

IV. The basic searching steps are as follow:

The search (query) image q

Step 1: Local features extraction from the input image (section 3.4).

Step 2: Find the bins by locating the closest feature centroids. The key values are represented by the number of bins.

Step 3: Find the specific two bins by using the global key to search the 2 B-trees simultaneously.

Step 4: Half-search to establish the individual features within the bins.

Step 5: Retrieve the whole candidates.

Step 6: Compare q with the retrieved candidates.

Step 7: Find the probable match.

3.6 Summary

Chapter 3 presents the methodology used in this study. The first part discussed the techniques of partitioning and clustering used. An enhanced feature extraction method was designed to extract the most discriminating information from the iris images in the form of sub-blocks. The multi-resolution analysis was used for each sub-block of iris image to produce the compact local image features. DWT was used to decompose the images into seven sub-bands, DCT represents the information in different frequency components of each sub-bands and SVD analysis was used to capture the local texture information and to generate a two dimensional 14 feature values which represent the most relevant features of each sub-blocks of image. Furthermore, an improved algorithm (WKIFA) was introduced for improving the clustering performance and convergence rate of algorithms. Finally, the searching technique based on a half-searching algorithm and a parallel search approach was presented.

CHAPTER 4

RESULTS AND DISCUSSION

4.1 Overview

In this chapter, the results of the experiments were discussed. The discussed results were in three segments: the first segment presented the efficiency and accuracy of the indexing method; the second segment analyzed the performance of the improved clustering method, and the third segment explained the complexity of the system. This chapter is structured as follows: Section 4.2 presented the experimental setup, including the types of iris database used in the system. Furthermore, this section analyzed the efficiency and accuracy of the indexing approach in terms of the penetration rate and bin miss rate. Section 4.3 compared the validity and feasibility of the proposed WKIFA based on the multi-peak function with the standard and existing methods. The effect of WKIFA on the experimental results of indexing method was also checked. Section 4.4 presented an analysis of the time complexity of the proposed searching and retrieving methods and compared them to the existing methods. Section 4.5 presented the summaries of the results and their discussion.

4.2 The Performance of the Indexing System

The performance of partitioning, clustering and indexing system is usually assessed based on their penetration rate (PR) for efficiency and bin miss rate (BM) for accuracy (Mehrotra 2010; Barbu & Luca, 2015; Dey & Samanta, 2012). Penetration rate which is represent the percentage of total database to be scanned on an average for each search, lower the penetration rate, more efficient the system and Bin miss rate which is obtained by counting the number of genuine biometric samples that have been miss-placed in a wrong bin (Kavati *et al.*,2017; Mehrotra, 2010) as higher accuracy and efficiency are measured by lower error rate and lower penetration rate (Kavati *et* *al.*,2017; Jiang, 2009; Mehrotra, 2010). A measure of these performance matrixes is dependent on the type of biometric database used. Hence, it is necessary to check the performance of different algorithms on the same database

4.2.1 The Efficiency of The system

The efficiency of the indexing system is assessed using Penetration rate (PR). the query feature set is compared to all the other templates in the database. Search efficiency can be achieved by partitioning the database based on some criteria. Thus, during identification, the query template is compared to only select templates in the appropriate partitions. The portion of total database to be scanned on an average for each search is called penetration coefficient PR, which can be defined by (Dey & Samanta, 2012 ; Kavati *et al.*,2017).

$$PR = \frac{1}{M}\sum_{i=1}^{M}\frac{C_i}{N} \qquad 4.1$$

where C_i represents the individual set of the i^{th} test image, N represents the total number of images in the dataset, and M represents the number of images to be scanned. A good indexing scheme will achieve a high hit rate (low BM) and a low penetration rate (PR).

4.2.2 The Accuracy of The System

Bin Miss Rate is used to assess the accuracy of the indexing system. A bin error occurs when an attempt is placed in a bin which is not compared with the correct bin for the biometric entity used, and hence will fail to match. The error occurs due to misplacing of biometric template in the wrong bin during identification (Kavati *et al.*, 2017).

$$Bin\ Miss\ Rate = 100 - Hit\ Rate \qquad 4.2$$

To analysis the efficiency of the enhanced approach, the samples were partitioned into galleries and test sets. The gallery set consists of 80% samples of all identities while the test set accounts for the remaining 20%. The gallery set samples were randomly selected, but for the test set, all the galley set samples were enrolled into the database to find the best match for each test set sample. The following processes were conducted during the experiments to compute these measures. Assume a database with *N* samples and *M* query samples. For the query sample q_i, C_i was defined as the least number of samples to be drawn from the database to guarantee a hit. Thus, the value (equation 4.1) will be the PR corresponding to a BM of (equation 4.2). Experiments were conducted using the mentioned databases and the results are as follows: The penetration rate (PR), bin miss rate (BM), and the retrieved features with different numbers of sub-bands for CASIAV3I, BATH, and IITK databases are presented in Table 4.1.

Table 4.1 Performance rates of the indexing method based on 2D features for change in the number of Retrieved features (RF) for different datasets

key	CASIA			BATH			IITK		
length	RF	BM	PR	RF	BM	PR	RF	BM	PR
1	4	0.0122	58.04	5	0.0301	23.21	5	0.0012	28.32
2	11	0.0128	20.41	17	0.1022	6.65	12	0.0400	12.00
3	28	0.0920	9.34	45	0.2318	2.59	31	0.0502	6.74
4	55	0.1231	7.22	98	0.3271	0.95	65	0.1066	2.87
5	98	0.2191	3.84	141	0.3391	0.61	97	0.1103	1.95
6	142	0.2700	2.28	241	0.3717	0.25	149	0.1412	1.01
7	199	0.3037	0.98	291	0.4226	0.13	276	0.2019	0.12

RF: Retrieved features, BM: Bin miss rate, PR: Penetration rate

An exact match the tree is traversed using all the sub-bands. However to obtain similar matches the tree traversal will stop before reaching the leaf and images having the same partial key is retrieved to find a match, The large set of images will be obtained using partial match which in turn increases the penetration rate. The bin miss rate and penetration rate is obtained by varying the number of sub-bands. With the change in the number of sub-bands the number of classes formed at node also changes. Table 4.1 shows the number of classes, penetration rate and bin miss rate by varying the number of sub-bands for CASIA, BATH and IITK databases. From the table it has been observed that with the change in the number of sub-bands the number of classes at node also changes. The number of classes increases when number of sub-band increased because with less number of sub-bands the length of global key reduces. The tree is not traversed completely till the node and the images that have same partial key are used to find the match. Hence probability of finding an image is higher in partial traversal compared to complete traversal. The bin miss rate reduces for partial traversal. However, partial traversal gives higher penetration rate due to increase in the number of templates stored in each class.

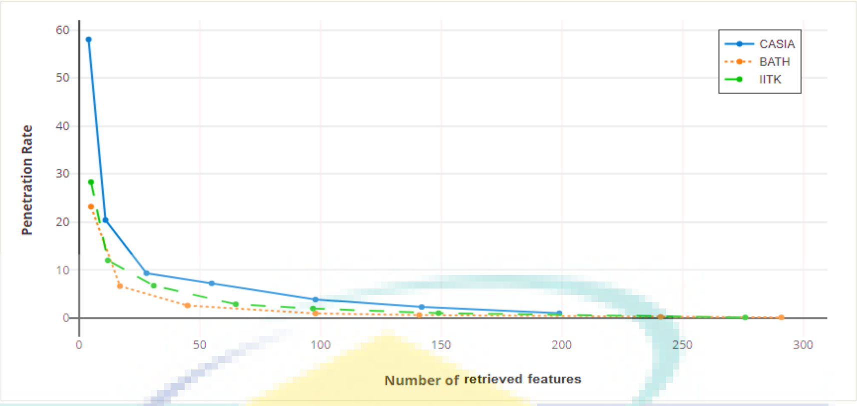


Figure 4.1 The PR when varying the number of features for different datasets

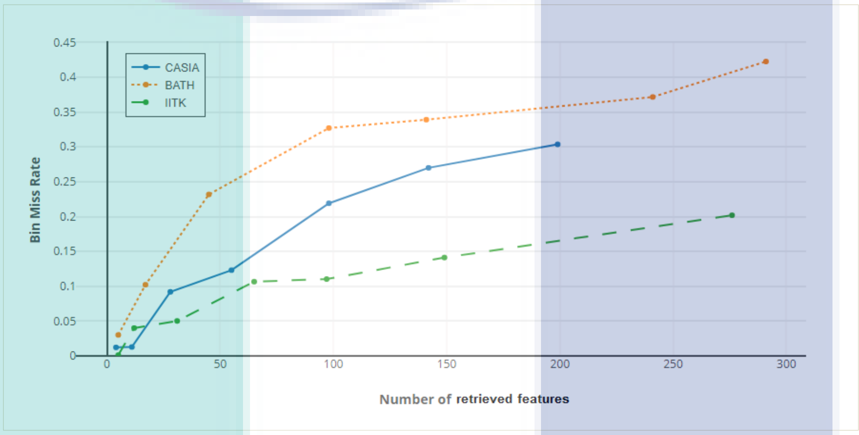


Figure 4.2 The BM when varying the number of features for different datasets

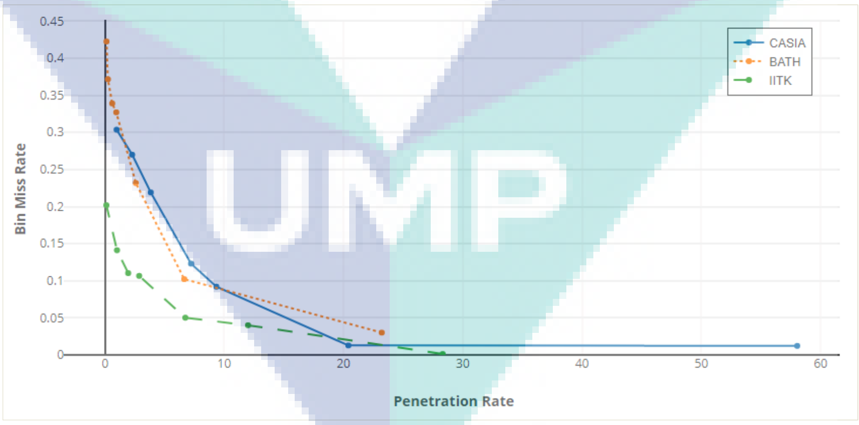


Figure 4.3 The PR-BM relationship for different datasets

Figure 4.1, Figure 4.2 show the Penetration Rate and Bin Miss Rate for change in number of features, the curves shows the difference between the three database CASIA V3I, BATH and IITK, where different iris databases give different measurements results. Figure 4.3 shows the relationship between PR versus BM for three databases CASIA V3I, BATH and IITK. In the existing method, only one value is applied to represent each iris images subband, which is inaccurate to express features especially for the subband 4×4 the S5, S6 and S7. For our indexing method, the first two singular values ($\sigma1$, $\sigma2$) from $\sum$ matrix are selected to represents each subband.

Table 4.2 Performance rates for indexing method when using 1D feature and 2D features for change in the number of Retrieved features (RF)

key length	1D feature vector			2D feature vector		
	RF	BM	PR	RF	BM	PR
1	4	0.0098	63.95	4	0.0122	58.04
2	13	0.2304	26.01	11	0.0128	20.41
3	39	0.3704	12.16	28	0.0920	9.34
4	74	0.5189	6.381	55	0.1231	7.22
5	127	0.5795	4.104	98	0.2191	3.84
6	193	0.6397	2.367	142	0.2700	2.28
7	276	0.7102	2.031	199	0.3037	0.98

RF: Retrieved features, BM: Bin miss rate, PR: Penetration rate

The following results show the evaluation of the indexing method in term of efficiency and accuracy when using one feature from each subband (one singular values) to create a one-dimensional features vector of (7 values). On the other hand, when extracting two features from each subband (two singular values) to create two-dimensional features vector of (14 values) to represent each of iris image. Table 4.2 shows the comparison of the performance of the indexing method when using one dimensional feature vector (7 features) and using two dimensional features vector (14 features) in term of BM: Bin Miss Rate in (%), PR: Penetration Rate in (%) and number of retrieved candidates for CASIAV3I. The indexing based on 2D features is outperforming to 1D feature because it is more efficient for representing the iris image's regions. With the change in the number of subbands the number of classes formed at node also changes. The table shows the number of classes, penetration rate and bin miss rate by varying the number of subbands. From the table it has been observed that with increase in the number of subbands the number of classes (#) also increases. This is because with less number of subbands the length of global key reduces. The tree is not traversed completely till the node and the images that have same partial key are used to find the match. Hence probability of finding an image is higher in partial traversal compared to complete traversal. The bin miss rate reduces for partial traversal. However, partial traversal gives higher penetration rate due to increase in the number of templates stored in each class. From the table when using two dimensional features vector if number of subbands is 1, CASIA database shows bin miss rate of 0.0122% and penetration rate of 58.04%. However if number of subbands is 10, the penetration rate reduces significantly to 0.98%.Thus there exists a tradeoff between the two evaluation rates.

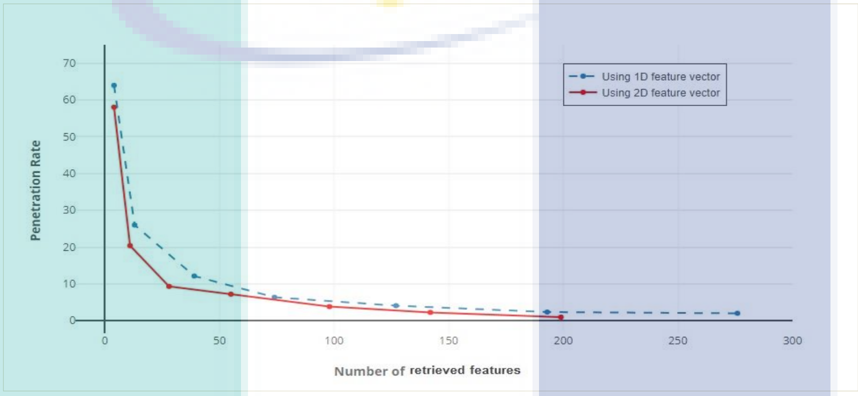


Figure 4.4 Comparison of the PR when both 1D and 2D features are used for CASIAV3I

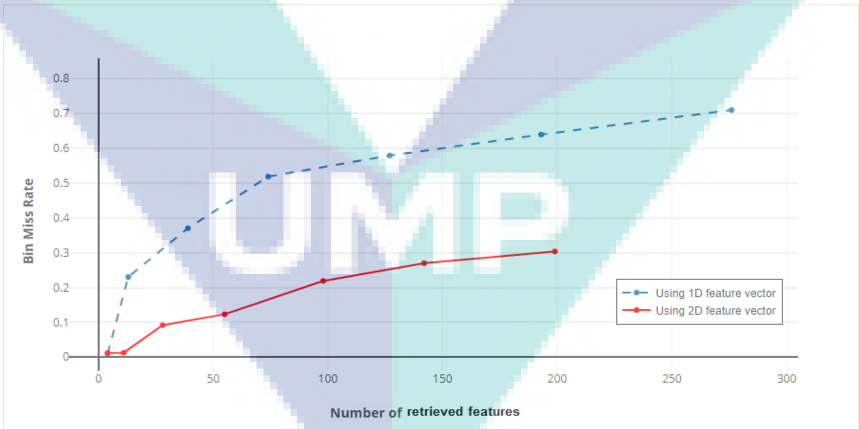


Figure 4.5 Comparison of the BM when both 1D and 2D features are used for CASIAV3I

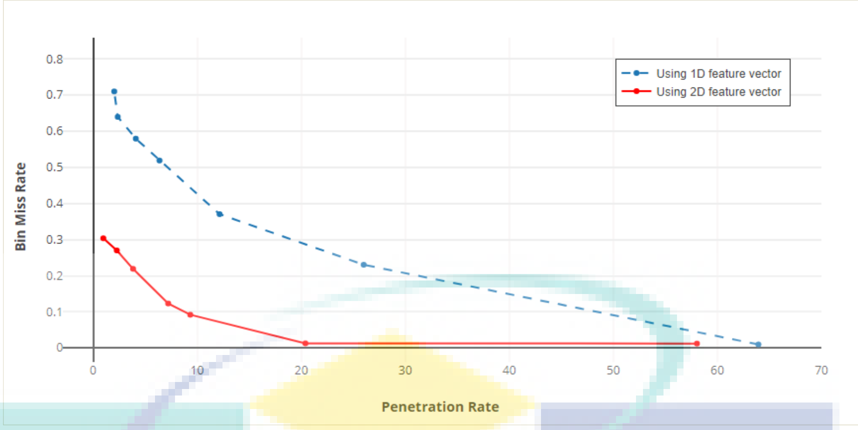


Figure 4.6 The PR-BM relationship for the two methods

Figure 4.4 depicts the changes in the Penetration Rate (PR) when changing the number of features. This figure was generated for the CASIAV3I database. Similarly, the Bin Miss Rate (BM) was plotted for different number of features Figure 4.5, while PR-BM relationship was plotted in Figure 4.6. The evaluation was also conducted using CASIAV4T database, Table 4.3 Presents the results of the indexing method using both one dimensional feature vector and two-dimensional feature vector with different class numbers, The results showed the efficiency of using two dimensional feature vector instead of one dimensional feature vector in term of efficiency and accuracy.

Table 4.3 Performance comparison when using 1D feature and 2D features for CASIAV4T

key length	1D feature vector			2D feature vector		
	RF	BM	PR	RF	BM	PR
1	7	0.0092	75.21	7	0.0021	65.32
2	83	0.1041	47.51	56	0.0851	35.71
3	157	0.3301	36.09	107	0.1430	13.22
4	243	0.497	15.20	196	0.3038	9.29
5	361	0.5438	9.21	231	0.4141	5.12
6	423	0.6926	5.37	297	0.5491	2.32
7	545	0.7581	2.67	350	0.6234	0.135

RF: Retrieved features , BM: Bin miss rate , PR: Penetration rate ,

Figures 4.7 and 4.8 shows the BM and PR for different numbers of classes on CASIAV4I database, Figure 4.9 showed the Penetration Rate versus Bin Miss Rate relationship for CASIAV4I.

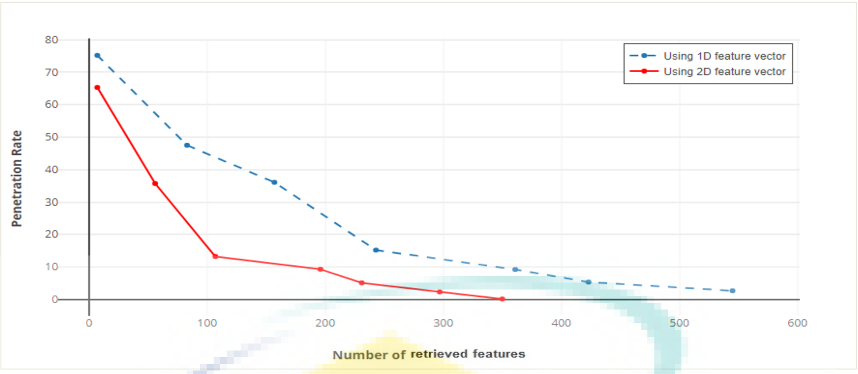


Figure 4.7 Comparison of the PR when both 1D and 2D features are used for CASIAV4I

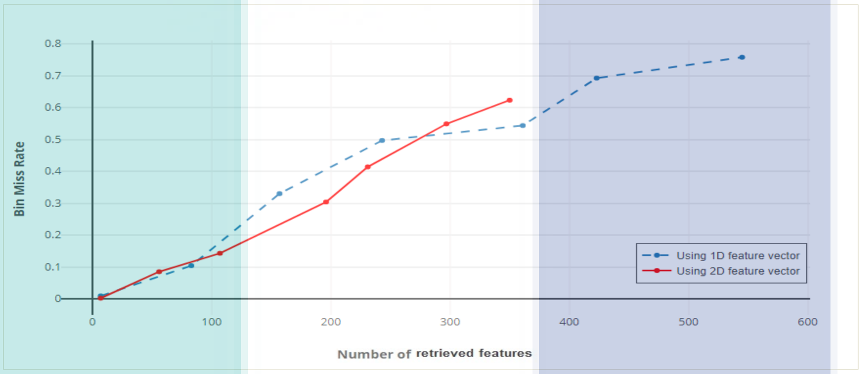


Figure 4.8 Comparison of the BM when both 1D and 2D features are used for CASIAV4T

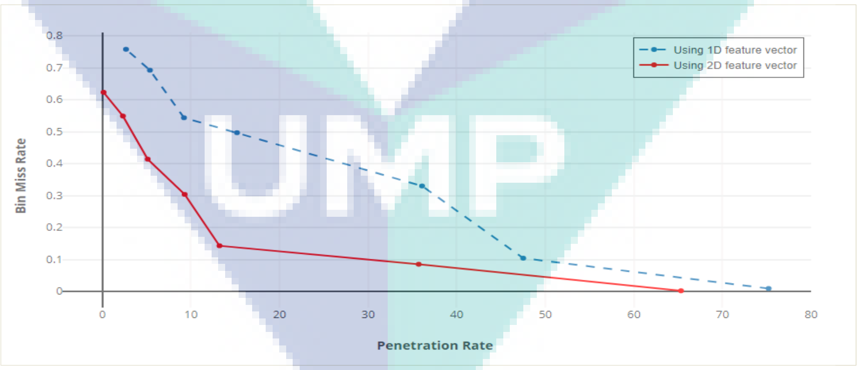


Figure 4.9 Comparison of the PR and BM when both 1D and 2D features are used for CASIAV4T

The method has been compared with two of the existing methods (Mehrotra *et al.*, 2009; Mehrotra, 2010) that use the same technique of indexing, while the researcher use different techniques for partitioning and clustering. These methods use only one value to express each subband which is the total energy value, where the multi-resolution subband coding of DCT coefficients is used to extract the energy features from the rectangular block, then forming an indexing key from the energy histogram. However, applying one value to express each subband is inaccurate, especially to express the 4×4 subbands of S_5, S_6 and S_7. In our improved indexing model, two features have been applied to express each subband, which are the first big two singular values ($\sigma1$, $\sigma2$) from $\sum$ matrix to represents each subband, where the singular vector expresses the feature for all subbands images. Because, the most important feature of an image can be expressed by the biggest singular vector values (Sadygov, 2014). The researchers decomposed iris images into ten subbands, each subband represented by only one feature value: the first method DCT Energy Histogram (Mehrotra *et al.*, 2009). The second method (Keypoint Descriptors) involves the improvement of the pre-processing phase to eliminate localized pupil boundaries and specular highlights before feature extraction.

Table 4.4 Comparison of the 2D features index method with the existing methods for CASIA V3I

key length	DCT Energy Histogram method			Keypoint Descriptors method			The Enhanced Indexing method		
	RF	BM	PR	RF	BM	PR	RF	BM	PR
1	2	0	99.69	4	0.1088	64.04	4	0.0122	58.04
2	5	0.016	35.69	13	0.2228	25.90	11	0.0128	20.41
3	16	0.036	22.70	39	0.3627	11.99	28	0.0920	9.34
4	39	0.132	10.23	73	0.5233	6.42	55	0.1231	7.22
5	82	0.24	6.12	129	0.5835	3.67	98	0.2191	3.84
6	158	0.308	3.46	198	0.6425	2.31	142	0.2700	2.28
7	233	0.356	2.63	272	0.7098	1.51	199	0.3037	0.98

RF: Retrieved features, BM: Bin miss rate, PR: Penetration rate

Table 4.4 summarizes the comparison results between our indexing method and the existing methods in term of Penetration Rate (PR) versus Bin Miss Rate (BM) using CASIA V3I database. The results show that our indexing method is more accurate and efficient than the existing methods, where the Penetration rate and Bin miss rate values is lower especially when fully index key is used as well as less number of candidates. Where lower the penetration rate, more efficient the system. In estimating penetration rate it is assumed that the search does not stop on finding the match but continues through the entire partition. However, using partial key mean less number of subbands and the tree is not traversed completely till the end of nodes and the images that have same partial key are used to find the match. Hence probability of finding an image is higher in partial traversal compared to complete traversal. Thus, for partial traversal the bin miss rate reduces and penetration rate increases. Experimental results show that our method outperforms existing methods, where the image is decompose into seven subband instead of ten subband as on the existing methods, moreover each subband is represented by two features instead one feature, besides the efficiency of the extraction approach for the image's features.

Figure 4.10 and Figure 4.11 showed an improvement in the efficiency and accuracy of most indexing results based on 2D features compared to the existing methods in term of PR and BM using different class numbers on CASIA V3I database. Figure 4.12 depicts the Penetration Rate versus Bin Miss Rate relationship for the three methods.

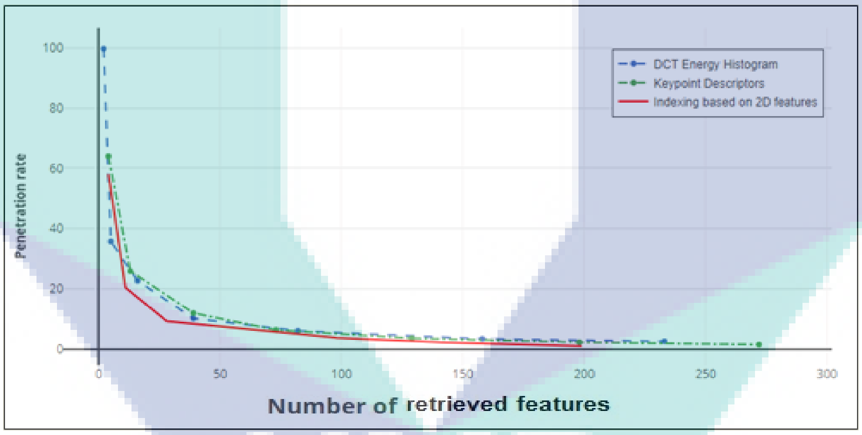


Figure 4.10 Comparison of the PR of the indexing method with the existing methods for CASIA

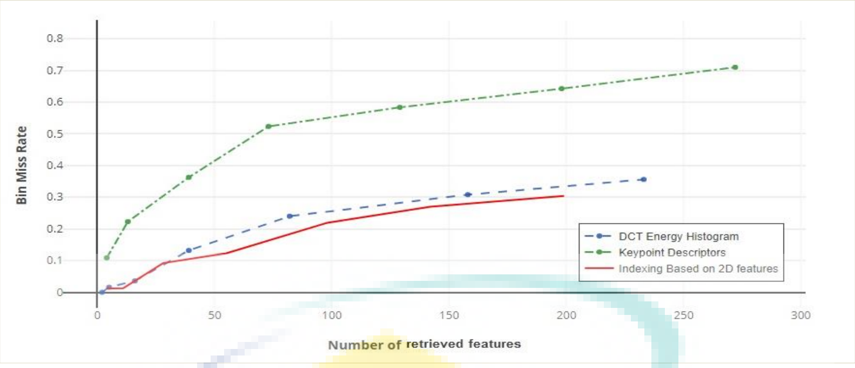


Figure 4.11 Comparison of the BM of the indexing method with the existing methods for CASIA

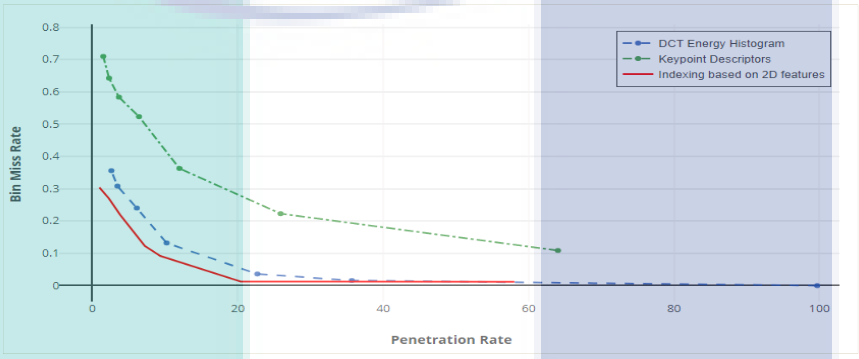


Figure 4.12 The PR-BM relationship for the three methods for CASIA

The experimental results of using BATH database are shown in Table 4.5. The efficiency and accuracy of the indexing method based on 2D features were better compared to the benchmarked methods in term of PR and BM when using different class numbers.

Table 4.5 Comparison of the 2D features index method with the existing methods for BATH

key length	DCT Energy Histogram method			Keypoint Descriptors method			The Enhanced Indexing method		
	RF	BM	PR	RF	BM	PR	RF	BM	PR
1	5	0.04	26.14	5	0.0444	26.14	5	0.0301	23.21
2	23	0.12	7.69	23	0.1333	7.69	17	0.1022	6.65
3	66	0.26	3.04	66	0.2889	3.04	45	0.2318	2.59
4	130	0.36	1.42	130	0.4000	1.42	98	0.3271	0.95
5	197	0.38	0.92	197	0.4222	0.92	141	0.3391	0.61
6	313	0.56	0.49	313	0.6222	0.49	241	0.3717	0.25
7	399	0.6	0.3	399	0.6667	0.3	291	0.4226	0.13

RF: Retrieved features, BM: Bin miss rate, PR: Penetration rate

Figures 4.13 and 4.14 showed the comparison of the indexing method based on 2D features with the current methods in term of PR and BM for different class numbers using BATH. The graph of the proposed method shows a better efficiency of the enhanced method compared to the other methods. Figure 4.15 showed the the Penetration Rate versus Bin Miss Rate relationship for the three methods.

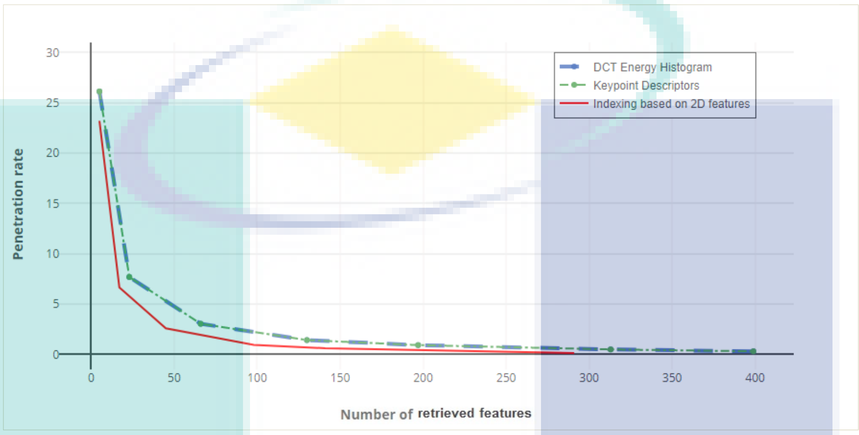


Figure 4.13 Comparison of the PR of the three methods for BATH

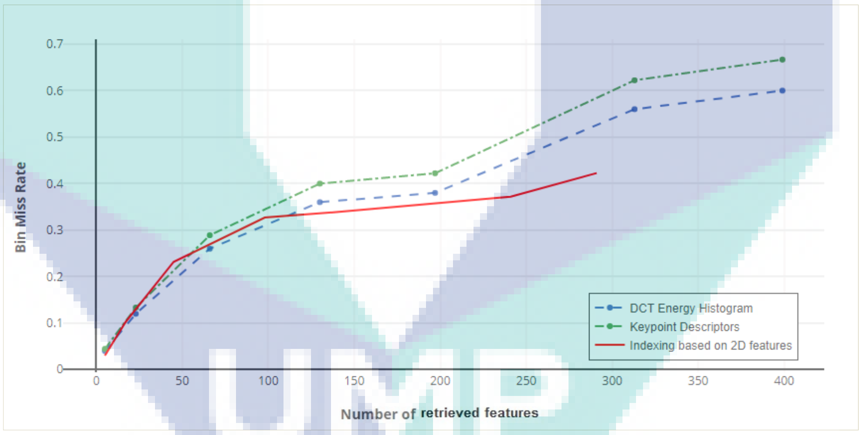


Figure 4.14 Comparison of the BM for the three methods for BATH

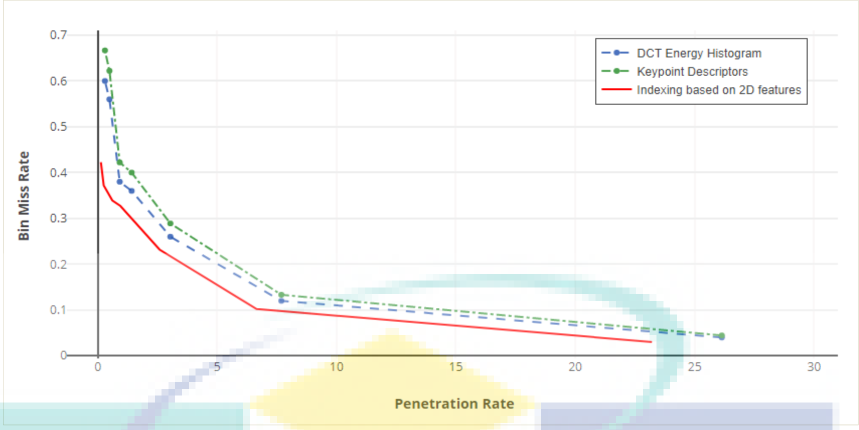


Figure 4.15 The PR-BM relationship for the three methods for BATH

The method was also tested using IITK database. The result in Table 4.6 showed an improve efficiency of the enhanced technique compared to existing techniques in term of the Penetration Rate versus Bin Miss Rate for different class numbers.

Table 4.6 Comparison of the 2D features index method with the existing methods for IITK

key length	DCT Energy Histogram method			Keypoint Descriptors method			The Enhanced Indexing method		
	RF	BM	PR	RF	BM	PR	RF	BM	PR
1	5	0.04	26.14	5	0.0197	41.44	5	0.0012	28.32
2	23	0.12	7.69	19	0.0658	17.21	12	0.0400	12.00
3	66	0.26	3.04	46	0.0724	09.24	31	0.0502	6.74
4	130	0.36	1.42	93	0.1316	4.77	65	0.1066	2.87
5	197	0.38	0.92	148	0.1645	3.25	97	0.1103	1.95
6	313	0.56	0.49	252	0.2039	1.56	149	0.1412	1.01
7	399	0.6	0.3	396	0.2697	0.92	276	0.2019	0.12

RF: Retrieved features, BM: Bin miss rate, PR: Penetration rate

The key is used for traversing the tree to arrive at the leaf node and retrieve the candidates. However, with increase in the length of key the number of candidates also increases. This is because the length of global key reduced, this mean the tree is not traversed completely till the leaf node and the images that have same partial key are used to find the match. So, if the complete key is used for traversing the tree then the probability of finding exact match becomes less. However, partial traversal gives higher penetration rate due to increase in the number of candidates. Figures 4.16 and 4.17 showed the efficiency of the enhanced method compared to the current methods in terms of PR and BM for different class numbers using IITK database. Figure 4.18 shows the PR-BM relationship for the three methods.

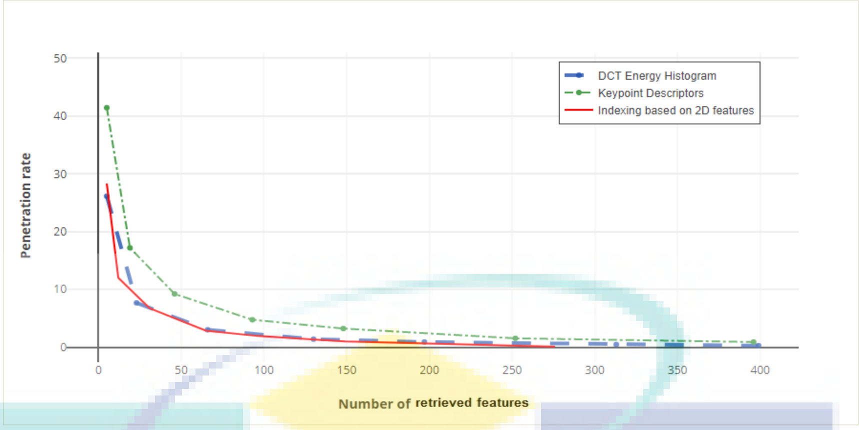


Figure 4.16 Comparison of the PR for the three methods for IITK

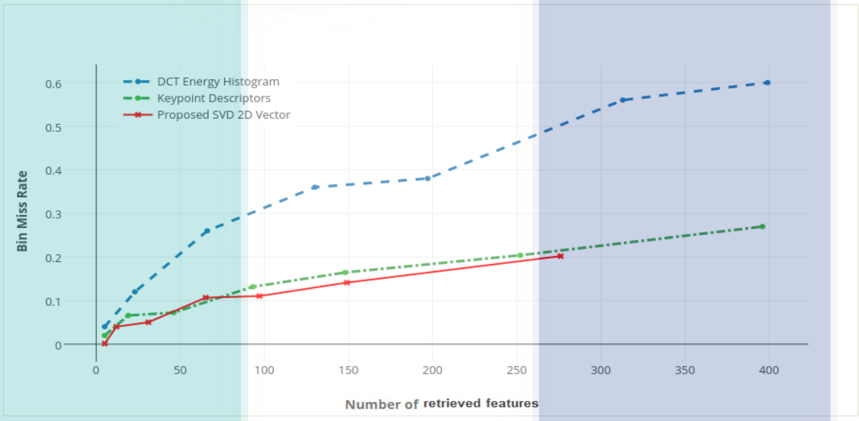


Figure 4.17 Comparison of the BM for the three methods for IITK

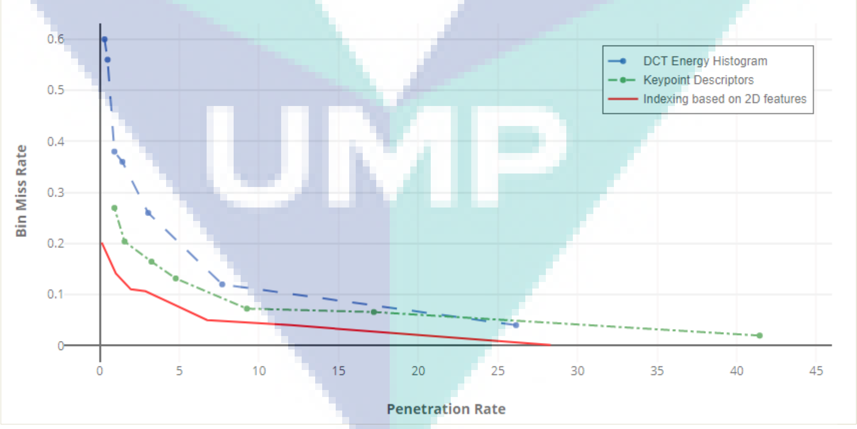


Figure 4.18 The PR-BM relationship for the three methods for IITK

Table 4.7 showed the comparison of the experimental results between the proposed method and the existing method in terms of PR. The represents the efficiency of the indexing method based on 2D features. Besides, the BM represents the accuracy of the indexing system.

Table 4.7 Comparison of the indexing method based on 2D features with the existing methods for CASIA V3I.

(Mehrotra et al method, 2010)		(Si et al method, 2012)		(Barbu and Luca method, 2015)		The Enhanced Indexing method	
BM	PR	BM	PR	BM	PR	BM	PR
4.15	5.43	2.62	2.408	0.16	3.482	0.0122	58.04
4.485	5.08	2.991	2.386	0.193	3.233	0.0128	20.41
4.775	4.97	3.508	2.32	0.386	2.124	0.0920	9.34
5.17	4.874	4.134	2.24	0.411	2.067	0.1231	7.22
5.56	4.806	5.25	2.216	0.493	1.831	0.2191	3.84
5.875	4.755	6.79	2.02	0.69	1.57	0.2700	2.28
6.316	4.62	8.22	1.89	0.83	1.13	0.3037	0.98

*Bold represents the best

The comparison of the developed method in this thesis with the other methods used in the literature showed the enhanced method based 2D vector to perform better than the benchmarking methods. This implies the fulfilment of the first problem identified in this study which is “the ambiguity of extracting important features from biometrics to distinguish between different biometrics and the relations between them using three popular transformation methods called DCT, DWT, and SVD”. The first objective of this study is to “improve the efficiency and accuracy of indexing technique by hybridizing three transformation methods (DCT, DWT, and SVD) to extract sufficient numbers and most relevant local features from iris images”. The current section covered the first objective. The next section will cover the second problem related to the thesis.

4.3 The Efficiency of WKIFA Algorithm

4.3.1 Experimental Data

To verify the validity and feasibility of the algorithm, the experiment mainly includes two parts: using the multi-peak function of (Yang, 2009) to test the validity and feasibility of the firefly algorithm in finding the initial clustering centre. The WKIFA algorithm was compared with the traditional K-means and FA algorithm on the standard UCI dataset based on simulation experiments. The advantages of these algorithms were compared based on the experimental results. The composition of the experimental dataset is shown in Table 4.8. For an easy and efficient operation, the UCI dataset was first standardized and normalized: the normalized Equation is as follows (using the mean):

$$x^* = \frac{x_i - \bar{x}_i}{S_i} \tag{4.3}$$

Where $\bar{x}_i$ = data mean, S_i data standard deviation

Table 4.8 Composition of the experimental sample data set

Dataset	Number of classes	Properties dimension	Dataset size
iris	3	4	150
wine	3	13	178
seed	3	7	210
Glass	6	9	214
Hayes-Roth	3	4	132
New-thyroid	3	5	215

4.3.2 Cluster Center Point Selection Experiment

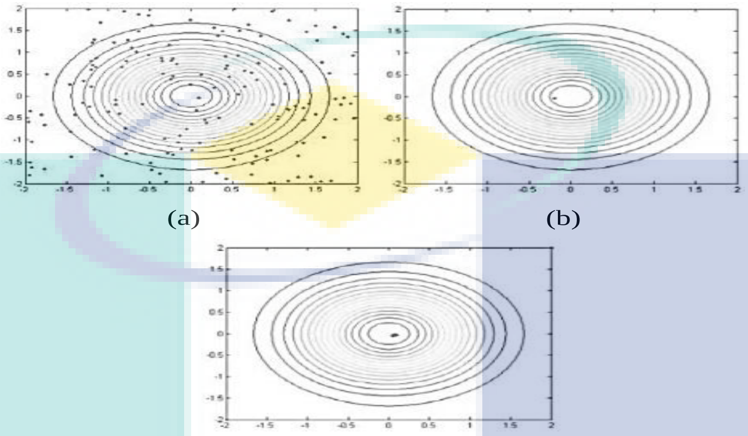
This section studied the evolution of algorithms over time under certain parameter settings that are necessary for developing an effective and precise algorithm. These include several settings of 4 parameters which are the number of fireflies (*n*), *β*, *γ*, and the maximum generation(*T*). This chapter highlights the influence of single parameter variations. This section particularly investigated the scenarios presented in Table 4.9 and Figure 4.19. Furthermore, each designed case was performed 20 times with 50 repetitions for all the runs. Based on the experiments, case S7 was used as the basis for the tests in this section. The parameters of this case were set at $\beta_0 = 1$, $\gamma = 1$, $\alpha = 0.06$ after repeated runs. The test results are shown in Figure 4.19.

Table 4.9 Convergence Scenarios

Scenarios	$\boldsymbol{\beta_0}$	***N***	γ	$\boldsymbol{T_{\max}}$
S1	0.9	1	0.1	100
S2	0.95	20	0.2	100
S3	0.98	40	0.4	100
S4	0.98	60	0.5	100
S5	0.99	120	0.7	100
S6	0.99	80	0.8	100
S7	**1**	**100**	**1**	**100**

*Bold represents the best

The reason behind chosen case S7 because this case can exploit the advantages of the Firefly search for a simultaneous scanning of several search regions while parallels spotting on the most significant regions, This will consequently result in achieving most of the most ideal solutions. Where the firefly search depends on the light absorption coefficient strength to achieve good solutions. Similarly, the Firefly search can explore the unexploited regions in a solution space using the light absorption coefficient (β = 1).



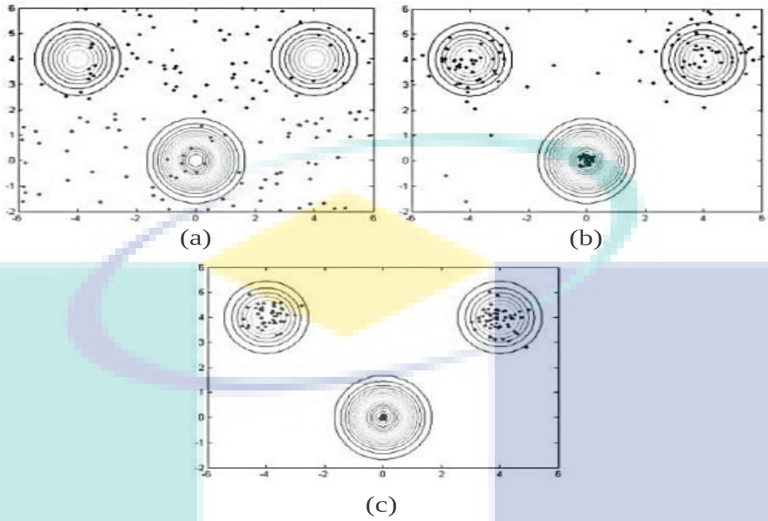
(a) (b)

(c)

Figure 4.19 Selection of the clustering center points

Figure 4.19 (a) showed the randomly selected 150 sample data in the *x*, *y* plane of the projection map. Figure 4.20 (b) showed the use of traditional FA algorithm through 4 iterations based on the selected 0.507s cluster center point. Figure 4.20 (c) showed that the algorithm through the two iterations, after 0.413s, obtained the clustering center point. Figure 4.20 showed that the clustering center point chosen by this algorithm was closer to the extreme point (peak point) of the peak function, and the selected firefly has the largest brightness and stability. It is shown that the calculation method, the disturbance mode, and the position updating equation of the firefly attractiveness can be improved more efficiently.

In order to verify the clustering center point chosen by this algorithm, the researcher can get a better clustering effect. Using the multimodal function in (Yang, 2009), set the appropriate range and select the three peaks to test. The same random selection of 150 sample data scattered in the solution space was used; the experimental parameter setting was the same as above, the clustering results are shown in Figure 4.20. In this figure, (a) showed a randomly selected 150 sample data in the x, y plane of the projection map, (b) showed the traditional FA algorithm-selected clustering center points, and (c) is a clustering result graph based on the clustering center point selected by the algorithm. After the algorithm has been improved, the sample points converged to the peak of the function, and the convergence effect was better. It is shown that the clustering center point of this algorithm can get more accurate and stable clustering.


(a)

(b)

(c)

Figure 4.20 Clustering results

4.3.3 Clustering Test Results and Analysis

Selection of standard data sets for simulation experiments: The experimental parameters were set as follows: Clustering - the number of classes $k = 3$, the number of fireflies = 100, $\beta_0 = 1$, $\gamma = 1$, $\alpha = 0.06$; maximum number of iterations $T_{\max} = 100$; $\beta = 0.9$; the iteration stopping threshold $\varepsilon = 0.105$, as previously shown. After the cluster center is updated, set the new border control in order to avoid the fireflies from being out of scope [range = -2.4, 3.1]. Weight coefficient m to take 2. To verify the accuracy of WKIFA and its feasibility, it was compared with the traditional K-means algorithm, FA algorithm in Iris, Wine, Seed, Glass, Hayes-Roth, and New-thyroid. Six different data sets were used; the average run time and an average number of iterations are shown in Tables 4.9, 4.10, and 4.11.

Table 4.10 Mean clustering results of the algorithm (%)

Dataset	K-means Algorithm	Yu, et al. Algorithm	FA algorithm	The improved WKIFA Algorithm
Iris	87.93	90.56	91.13	92.16
Wine	56.85	70.23	71.36	72.15
Seed	86.97	88.07	88.89	90.46
Glass	54.05	57.18	57.46	63.12
Hayes-Roth	77.32	81.06	79.25	82.35
New-thyroid	72.34	79.63	78.12	80.28

*The bold is the best

Table 4.11 The average iteration time of the algorithm (s)

Dataset	K-means Algorithm	Yu, et al. Algorithm	FA algorithm	The improved WKIFA Algorithm
Iris	0,044	0.038	0.033	0.025
Wine	0.048	0.043	0.038	0.037
Seed	0.056	0.053	0.053	0.049
Glass	0.103	0.093	0.092	0.089
Hayes-Roth	0.092	0.083	0.087	0.083
New-thyroid	0.088	0.085	0.092	0.081

*The bold is the best

Table 4.12 Average number of iterations of the algorithm

Dataset	K-means Algorithm	Yu, et al. Algorithm	FA Algorithm	The improved WKIFA Algorithm
Iris	7	4	3	3
Wine	10	9	8	7
Seed	9	7	7	6
Glass	11	9	10	7
Hayes-Roth	9	6	8	5
New-thyroid	10	7	8	6

*The bold is the best

The convergence curves of WKIFA, K-means (traditional), FA, and Yu et al. algorithm in Iris, Seed, Glass datasets are shown in Figure 4.21.

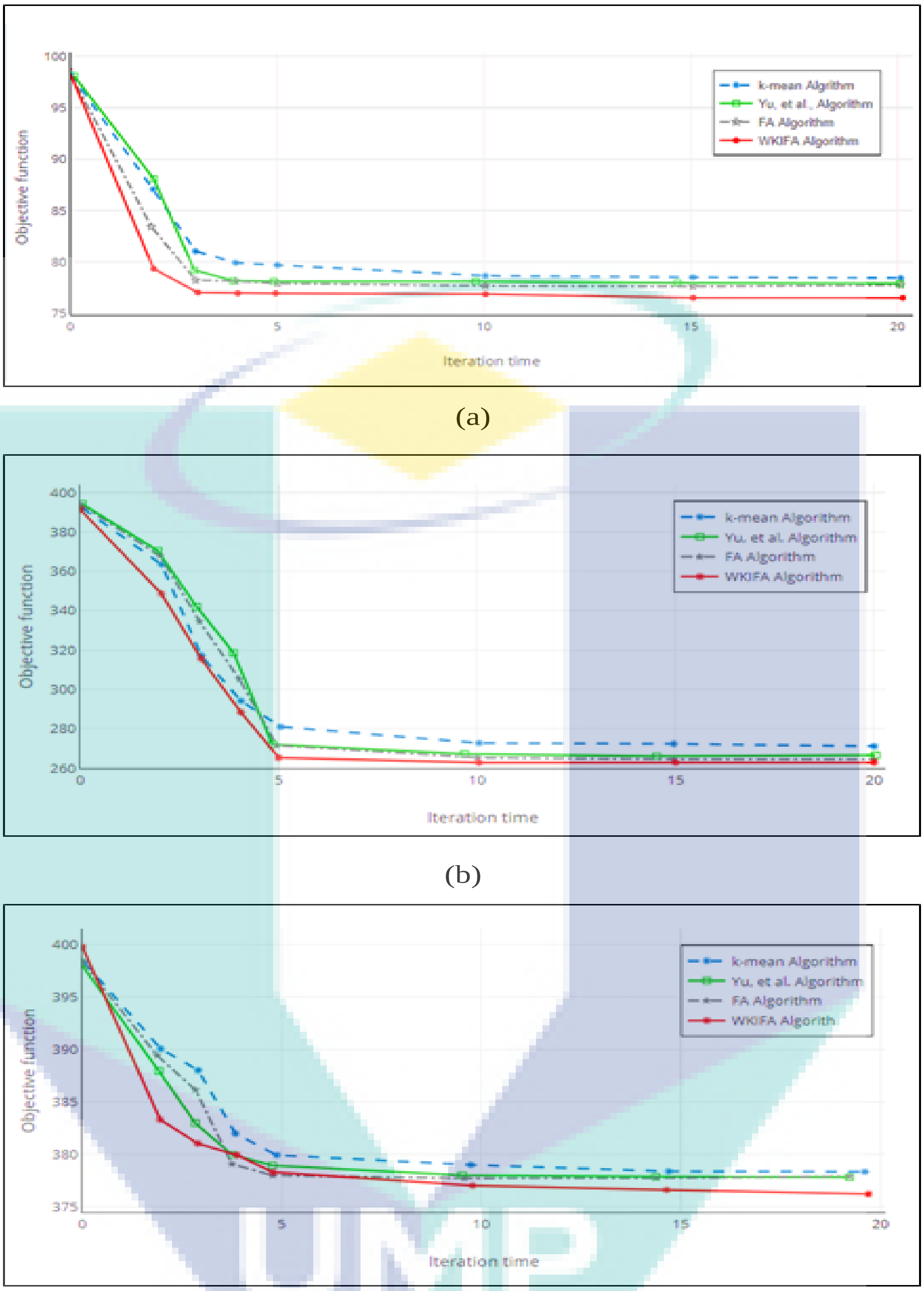


Figure 4.21 Comparison of convergence rates of the four algorithms on a) Iris, b) Seed, and c) Glass.

From the simulation results in Tables 4.10, 4.11, and 4.12, WKIFA algorithm showed a higher clustering accuracy compared to the other methods. Figure 4.21 showed a high clustering accuracy by improving the algorithmic convergence rate. WKIFA algorithm reached to the highest accuracy (lower objective function) with the minimum iteration compared to the other algorithms. Since this algorithm introduced the weighted Euclidean distance, the corresponding objective function (fitness function) is also changed accordingly. So, the original distribution is not obvious with the other algorithms such as the traditional K-means, FA, and Yu, et al. algorithm.

4.3.4 The Results and Analysis of the Indexing System Using WKIFA

WKIFA was used as an indexing system for automatic partitioning and clustering tasks instead of the existing traditional K-means. The final results were evaluated and compared with the existing results. The comparison of the final results of the indexing system using traditional k-mean algorithm and WKIFA algorithm in term of BM and PR were shown in Tables 4.13, 4.14, and 4.15.

Table 4.13 The performance of the system using WKIFA algorithm for BATH database

key length	Indexing method using K-means Algorithm			Indexing method using the improved WKIFA Algorithm		
	RF	**BM**	**PR**	**RF**	**BM**	**PR**
1	5	0.0301	23.21	5	**0.029**	**21.816**
2	17	0.1022	6.65	13	**0.0842**	**4.29**
3	45	0.2318	2.59	43	**0.2033**	**1.89**
4	98	0.3271	0.95	86	**0.2811**	**0.722**
5	141	0.3391	0.61	122	**0.2981**	**0.58**
6	241	0.3717	0.25	206	**0.301**	**0.206**
7	291	0.4226	0.13	218	**0.309**	**0.088**

* **RF**: Retrieved features, **BM:** Bin miss rate, **PR**: Penetration rate

Table 4.14 The performance of the system using WKIFA algorithm for CASIA database

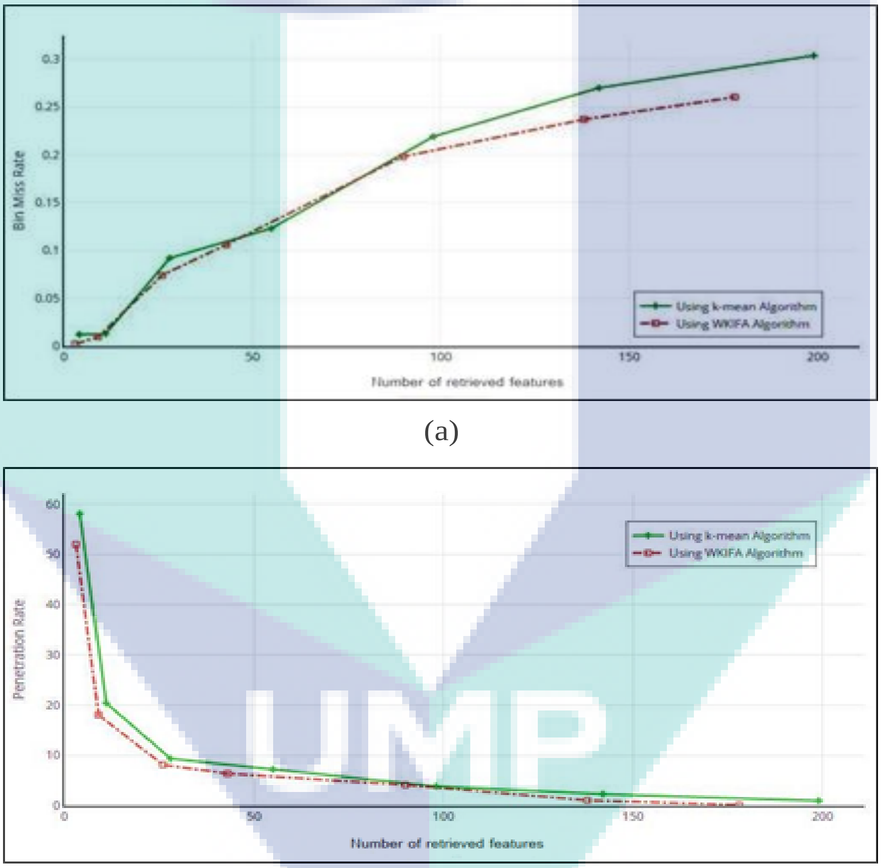
key length	Indexing method using K-means Algorithm			Indexing method using the improved WKIFA Algorithm		
	RF	**BM**	**RF**	**Class**	**RF**	**PR**
1	4	0.0122	58.04	3	**0.00231**	**51.95**
2	11	0.0128	20.41	9	**0.0097**	**18.01**
3	28	0.0920	9.34	26	**0.0742**	**8.107**
4	55	0.1231	7.22	43	**0.106**	**6.381**
5	98	0.2191	3.84	90	**0.1981**	**4.104**
6	142	0.2700	2.28	138	**0.237**	**1.067**
7	199	0.3037	0.98	178	**0.2604**	**0.131**

* **RF**: Retrieved features, **BM:** Bin miss rate, **PR**: Penetration rate

Table 4.15 The performance of the system using WKIFA algorithm for IITK database

key length	Indexing method using K-means Algorithm			Indexing method using the improved WKIFA Algorithm		
	RF	BM	PR	RF	BM	PR
1	5	0.0012	28.32	4	**0.0011**	**23.87**
2	12	0.0400	12.00	10	**0.0212**	**8.22**
3	31	0.0502	6.74	29	**0.0283**	**4.862**
4	65	0.1066	2.87	58	**0.1009**	**2.722**
5	97	0.1103	1.95	89	**0.1081**	**1.508**
6	149	0.1412	1.01	136	**0.122**	**1.009**
7	276	0.2019	0.12	213	**0.1548**	**0.108**

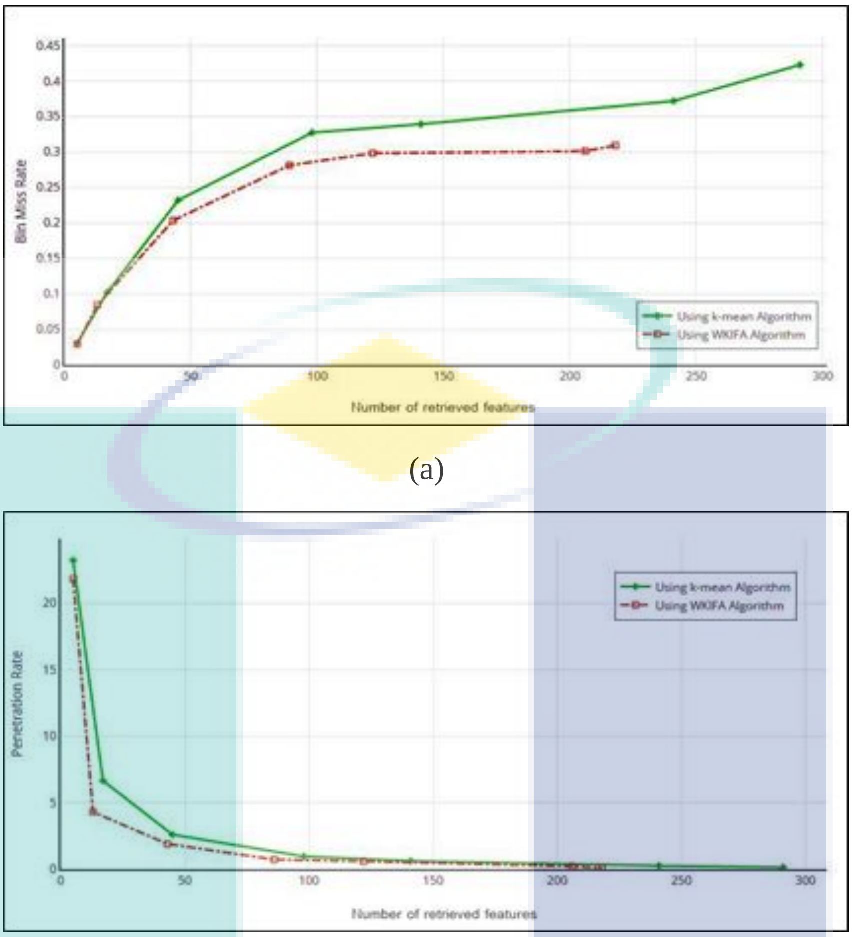
* **RF**: Retrieved features, **BM:** Bin miss rate, **PR**: Penetration rate



(a)

(b)

Figure 4.22 Comparison of the performance of the systems using the two algorithms for CASIA in terms of a) BM, and b) PR

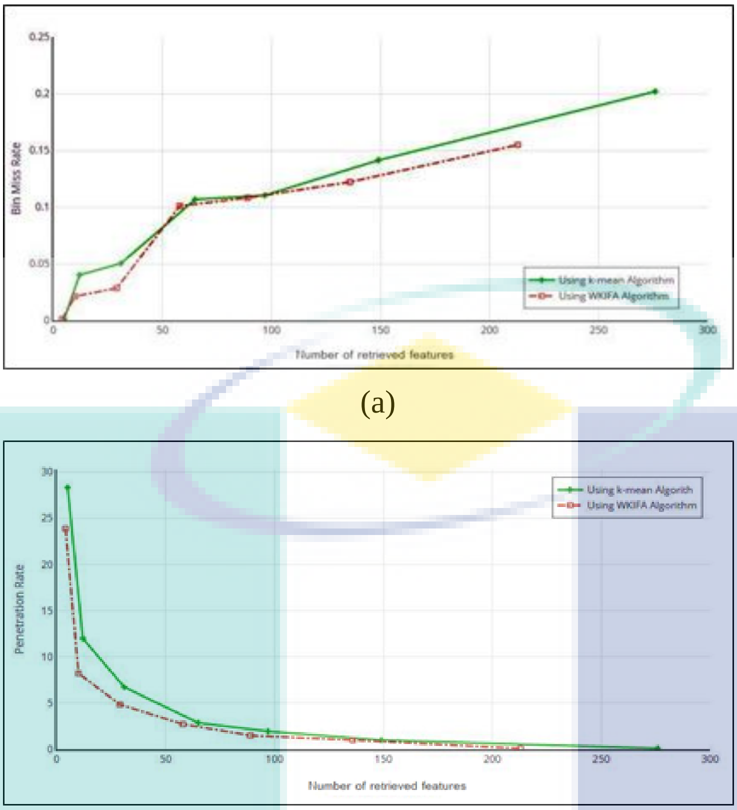


(a)

(b)

Figure 4.23 Comparison of the performance of the systems using the two algorithms for BATH in terms of a) BM, and b) PR

The results presented in Figures 4.21 to 4.24 showed a better clustering performance of WKIFA compared to the traditional k-mean. The improved performance was evidenced in the reduced PR of the system in the 3 databases (CASIA = 0.131, BATH = 0.088 and IITK = 0.108), as well as an improved accuracy as evidenced in the reduced BM (CASIA = 0.2604, BATH = 0.309, and IITK = 0.1548). The initial clustering centers of traditional K-means algorithm were optimized using the FA. At the same time, the defects due to noisy data and other uncertainties were reduced using a kind of weighted Euclidean distance which led to improved clustering processes.



(a)

(b)

Figure 4.24 Comparison of the performance of the system using the two algorithms for IITK in terms of a) BM, and b) PR

Besides these standard databases which have been mentioned in section 4.2.1, another new four training sets of iris images were created randomly from the four used databases with different sizes, the experimental results using these four datasets are shown in the Appendix A (Table A.1) The experimental results showing the efficiency of the enhanced approach.

4.4 Complexity Analysis of Retrieval

The researcher analyze the retrieval efficiency by measuring the average time taken to retrieve iris templates from the database for a given query iris, time is often expressed by Big O notation ("Big O notation", 2017). Let *tp* be the average time to perform addition, subtraction, and assignment operations. The indexing approaches require *N* comparisons to retrieve candidates corresponding to one feature and a candidate set of size *IL* is retrieved using 7 features. Therefore, the time taken for retrieval a candidate set of size *IL* is: ($tp \times \text{Log}(N)) \times 7$. Thus, our indexing approach takes less time than the linear search approach because $IL << N$.

Comparing with the existing methods to retrieve set candidates of size IL corresponding to all features: Dey & Samanta, (2012) developed a method based on the Gabor energy from different scales and orientations of the iris image the researchers generated a 12-dimensional index key texture features. The values of index keys of all individuals were used to create the index space. From the index space and the index key of the query image, since, the researchers use 12 key length the time taken for retrieval a candidate set of size *IL* is: $(tp \times N) \times 12$ .Mehrotra, (2010), which is based on energy histogram derived from iris texture, features are extracted after transfer iris image using DCT method and divided the transferred image into 10 subband, one value represented each subband which is the total energy value. Normally, the image is stored in the database together with 10 key length during enrolment. The extracted feature values of all individuals are uses to create the index key, for the query image, the researcher use 10 key length the time taken for retrieval a candidate set of size *IL* is: $(tp \times N) \times 10$

The researcher conclude that the enhanced method which is retrieving the query image using 7 features only, as well as using half search algorithm instead of liner search inside the created group, made the time taken for retrieval a candidate set of size *IL* is: $(tp \times Log (N)) \times 7$, Which less than the existing methods and the linear method. The retrieval execution times (specific to a computing environment of Intel Core i7-2600K Quad-Core Processor, 3.4 GHz, 8 MB Cache, 8.0-GB Memory) of the enhanced method determine the closest matching cluster according to the cluster centers comparison. In the specified computing environment, the retrieval execution time of the enhanced method was 0.0025 ms. In this method, the tree structure was utilized to save the iris data because the tree structure is easy to be extended. In the tree structure, the table was used to describe the relationship among elements. Compared with the linear storage of (Dey & Samanta, 2012), in the new table structure, one more column design was used to save the relationship. When the new data will be inserted, the relationship can be described by the tree structure expendable.

In addition, the storage space occupied by an algorithm on the computer memory includes the storage space occupied by the storage algorithm itself, the storage space occupied by the input and output data of the new algorithm, and the storage space that the algorithm takes temporarily during the operation. The algorithm occupies the storage space that is temporarily occupied during the running process. In the enahnced method, the space complexity of updated operations is $O(1)$ because the updated operation only need one more temporarily occupied space to save the new data in each time.

In the Dey & Samanta method, Gabor transform is simply the design of Gabor filter, and the design of the filter is the design of its frequency function (U, V) and Gauss function parameters (one). In fact, the Gabor transform is used to extract the local information of the signal Fourier transform. A Gauss function is used as the window function. Since the Fourier transform of a Gaussian function is a Gaussian function, Fourier inverse is also local. In addition, Gabor transform can select many texture features, but Gabor is non-orthogonal, and there is redundancy between different feature components; so, the efficiency is not too high in the analysis of texture images. However, the orthogonal transforms were used in the enhanced method. The iris image features can be extracted by extracting the local information, therefore, the accuracy is clearly improved.

In Rathgeb et al., (2015) method, the bloom filters were deployed to test the membership of a set. Elements can be introduced into the set, but cannot be removed. As more elements are introduced to the set, the chances of false positives increase. In the retrieval process, the B-tree is used to search for the elements. In each time, the retrieval process will rely on the depth of the tree. Because a binary vector was used for representing the elements, the retrieval time relied on database size.

4.5 Summary

This chapter presents the results of the experiments and their discussions using the three methods. Firstly, the results of the partitioning and clustering of iris images using the combined transformation methods were discussed. The enhanced method based on the experimental results showed efficiency in reducing the search space and high accuracy by reducing the number of candidates and reducing the BM. Hence, the combined transformation method filled the first gap highlighted in this study and satisfies the first objective of the study. The validity and feasibility of the improved algorithm (WKIFA) were also tested on achieving good clustering performance and convergence. The clustered experiments and validity tests were conducted on several groups of UCI data. The results showed great efficiency and superiority of the improved algorithm. Finally, the enhancement of the retrieval approach was done by separating each group of features into two groups and forming two B-trees. The searching was done in parallel and based on a Half searching algorithm within each group. The generated indexing keys were used to traverse the tree so as to reach the specific groups. On the other hand, half searching method was used to reach the candidate features. This method showed efficiency in candidate's retrieval by reducing the complexity and execution time. Therefore, this helped to solve the second problem identified in this study and satisfies the second, third, and fourth objectives.

CHAPTER 5

CONCLUSION

5.1 Introduction

This chapter concludes the work reported in this thesis. A brief summary and discussion of the study findings is presented in the subsequent section followed by some recommendations for future studies.

5.2 Study Findings and Discussion

Biometrics is yet not a full proof method of automatic human recognition, in spite of the fact that it appears to be the obvious technology for robust personal identification. Currently, inexpensive and compact biometric sensors, as well as fast processing chips are available and with these it is becoming increasingly clear that a broader use of biometric technology would provide better results. But there are three fundamental barriers which need to be overcome. These include accuracy, scalability and slow response which all contribute to the main challenges faced in the implementation of a biometric system. Normally, an ideal biometric system should be highly accurate, with a quick response potential, convenient to use, and easily scalable to a large population. With reference to these, the major obstacles that hinder the design and development of such a system are currently the thrust areas of research.

This study shows the importance of clustering and indexing for retrieval process and vice versa. As earlier stated, the indexing process arranges the information required for retrieval process, wherein if the image features are inappropriately represented, the goal of the retrieval process will not be achieved. On the other hand, a system that performs well with a particular database does not have 100% certainty to perform well with other different database. So, it is clear that indexing is affected by retrieval and images in the database. Since there is no universal indexing system that can handle every types of database, it is necessary to create a system that is as robust as much as possible. This study aimed to provide a better understanding of iris modality to improve the efficiency of retrieving processes by addressing the main problems associated with clustering, indexing and searching.

The first objective of this study has been achieved by improving the efficiency and accuracy of clustering and indexing biometric database, where a sufficient number of the most relevant local features of iris image have been extracted. This has been achieved through an enhanced features extraction technique. The iris image was initially divided into 8 x 8 blocks. From experiments, the best average identification was achieved when using images of size 8 x 8 compared to other image sizes. This process isolated the local features from the images. The analysis of small blocks makes transformation processes interesting and useful for image processing. Each block was transformed by the DCT and DWT transformation methods. A combination of the properties of DCT and DWT offered features extraction advantages such as the possibility of separating the information signals into low-frequency constituents with no correlation with their counterpart frequency index. It should be noted that the existing methods uses single frequency analysis which is not enough to get the holistic information of any iris pose variation. In another word, the DCT still have few missing dominant information of iris images. By this hybrid method, the space domain of image blocks has been transferred into a frequency domain expression which helps in the separation of the image into sub-bands of varying significance. The SVD transformation was used in this enhanced technique to analyze and decompose the regions of each sub-band into three simple transformations. This was followed by a subsequent selection of the first two singular values from the Σ matrix of each sub-band to represent local features. The most significant image features are represented by the biggest singular vector values. The selected block features are later combined to generate the features of an iris image. The features of each sub-band of all images were partitioned logically based on similarities. As such, images having similar texture patterns are placed together in the same group to have more accurate matches. In the case of partitioning-based clustering, the scalable K-means++ (K-means||) algorithm was selected based on its peculiar advantages such as speed and scalability.

The second objective of improving the clustering accuracy has also been achieved by presenting an enhanced clustering algorithm named WKIFA, to solve the drawbacks of k-means algorithm of noise sensitivity and outlier points. Likewise, it is well known that the K-means is highly sensitive to the selection of initial centres that may converge to the local optima of the criterion function rather than the global optima (Raykov *et al.,* 2016). The WKIFA was improved to overcome the mentioned shortages of k-mean by using Firefly Algorithm which has the potential for quick convergence rate and global search to optimize the initial clustering centers of the k-means algorithm. A weighted Euclidean distance was used to reduce the defects due to noisy data and other uncertainties. The performance of the k-means clustering method using either Euclidean distance function or Manhattan distance function was compared based on several parameters such as the number of iterations, the sum of squared errors, and the models' building time. From the analysis of the results, besides the mentioned advantages in the thesis, it was found that Euclidean distance outperformed the other distance functions when used with k-means Algorithm, where k-means clustering was analyzed with various distance functions such as Euclidean distance function and Manhattan distance function using the implementation 'WEKA' product and using dummy data. Likewise, the performance of the k-means clustering method using the two distance function was compared based on several parameters such as number of iterations, sum squared errors and time taken to build the model. From the analysis of the results, besides the mentioned advantages in the thesis, it was found that Euclidean distance outperformed the other distance functions.

The last objective is related to the data retrieval time and this has been achieved by introducing three iris modality-based methods for improving the efficiency of retrieving processes. These include the B-tree structure, parallel searching and half-searching algorithm. The half-searching algorithm was evaluated by analyzing and calculating the average case running time of a half search on n sorted elements and comparing it with the existing methods. It was found that the best-case running time of a half-search searching of n elements was O(1), corresponding to when the element of interest was found. For the linear search used in the existing methods, the worst-case run-time is O(n) and the average run-time is O(n). For enhancing the searching and retrieving processes of biometric systems, this research introduced an efficient searching approach which is initiated by ranking the features of each group, with a subsequent separation and distribution into two Bins of features to form two B-trees with specific global keys. The keys represent the group number it belongs. For the searching process, the public key is used to search the B-trees to find the required Bins. Finally, the half-searching method was used instead of the existing methods to find the candidates inside each Bin.

The system developed herein has been tested and evaluated on publicly available databases wherein the experimental results showed highly desirable efficiency of the indexing system. This was evidenced by the considerably low penetration rate 0.98%, 0.13% and 0.12% and lower bin miss rate of 0.3037%, 0.4226% and 0.2019% compared to the existing schemes for CASIA, BATH and IITK iris databases respectively. This is considered highly desirable because an efficient clustering and indexing method should have low miss rate and low penetration rate. The results of the improved clustering algorithm WKIFA showed that the improved method is more effective for clustering stage of the system and outperforms to the traditional k-mean, after the algorithm is improved. Interestingly, the sample points converge to the peak of the function, and the convergence effect is better. It is shown that the clustering center point of this algorithm can get more accurate and stable clustering besides saving the number of iteration and time. Thus, the performance of the indexing system has been well enhanced as evidenced through the Penetration Rate which was reduced to 0.131%, 0.088% and 0.108% for the mentioned database. Likewise, a better accuracy as evidenced through the Bin Miss Rate which was reduced to 0.2604%, 0.309% and 0.1548% for the mentioned database was also observed. On the other hand, analysis of time complexity of retrieval showed that the computational complexity is reduced to O (log N) which is better than the existing methods.

5.3 Research Contributions

Biometrics authentication (or realistic authentication) is used in computer science as a form of identification and access control. This study has made major contributions to the literature on biometrics, in addition to the provision of some directions for future research. This thesis adds value to research and practice communities concerned with biometric technologies. The contributions made throughout this research are diverse covering theoretical and practical facets.

5.3.1 Contributions to Theory

This dissertation theoretically contributes to knowledge by filling a gap in the literature regarding the indexing and retrieval of iris biometric data and highlight to the factors influencing the response time in large biometric applications. The theoretical contribution of this dissertation is a development of a substantive theory based on theory researches for the adoption of biometric authentication in large applications. For example: in the United Arab Emirates (UAE) which has launched a national border crossing security initiative, the system presently has 27 land, air and sea ports of entry which are equipped with the system (Mehrotra & Majhi, 2013). In India, a large scale project Aadhaar is mandated to issue unique identification number to each individual using fingerprint and iris (Kavati *et al.*, 2017). In the UK, the IRIS recognition system (IRIS) is used to enter through automated barriers at certain airports, these biometric systems face the scalability issue as the number of people to be rolled into the system runs into billions. This issue has become the bottleneck for low response time, high search and return efficiency in addition to accuracy (Dey & Samanta, 2012). The developed theoretical framework through this dissertation is unique and more comprehensive than other related existing theories.

5.3.2 Contributions to Practical

The biometric retrieval is a challenging task as the size of the biometric databases has increased considerably and there is no natural order by which one can sort the biometric data. On the other hand, the existing indexing and retrieval techniques only work well with small sized databases especially with respect to accuracy and time without providing a complete solution for larger databases. This thesis practically addresses this challenge by presenting an enhanced features extraction technique for iris image and an improved clustering algorithm for better partitioning accuracy. This was coupled with an efficient searching approach to improve the retrieval time of biometric system. The detailed of the practical contributions of this thesis are summarized as follows:

i. The thesis proffers more possibility for extracting sufficient and efficient number of local features of iris image in order to achieve an efficient clustering and indexing for biometric database. This was achieved by exploiting the synergized advantages of the three popular transformation methods: The Discrete Cosine Transformation (DCT), the Discrete Wavelet Transform (DWT), and the Singular value decomposition (SVD). The DCT and DWT were able to separate information signal into low frequency components uncorrelated with their counterpart frequency index. The DCT coefficients provide frequency component information but without localizing the specific frequencies in space. Interestingly, the DWT solves this problem by analysing the signal at different frequencies and at different times. On the other hand, SVD helped to reduce the amount of features and brings out the useful part of the data. This is because SVD can restructure a high-dimensional matrix into a low-dimensional matrix by eliminating the less important parts, thereby producing an approximate representation with any desired number of dimensions (rank).

ii. The thesis helps to improve the clustering performance of biometrics database, by reducing the impact of abnormal points while optimizing the initial clustering centers of K-means algorithm. This was achieved by presenting an enhanced clustering algorithm (WKIFA). This is based on Firefly Algorithm that has power ability of quick convergence rate and global search to optimize the initial clustering centers of the k-means algorithm. A weighted Euclidean distance was also used to reduce the defects related to noisy data and other uncertainties.

iii. The thesis offers a means of reducing the effect of increasing the number of features on the searching and retrieval time, by presenting an improved approach based on two B-tree structure, parallel searching and half-searching algorithm. In this approach, the groups of features are divided to form two b-trees based on the index keys. The generated keys were used to traverse the two trees in order to reach the specific groups whereas the Half searching method was used to reach the candidate features by searching within each group, thereby improving the response time for data retrieval.

5.4 Future Work

An extension of this study is recommended for subsequent research work especially with respect to image scaling and rotations. There is also a need to improve the clustering process by equalizing the size of the groups to make searching time for each image approximately constant. Moreover, the searching and retrieval processes can be enhanced by using a parallel B-tree construction method based on the Hadoop MapReduce framework. It is well known that B-tree usually takes a long time to construct especially for a huge data volume. Hence, it can be partitioned to build local B-trees in parallel before merging them to form a B-tree that covers the whole data set. On the other hand, the system can be extended by using multi-modal biometrics instead of a single one. This will increase the accuracy. Besides, when using fusion at the feature level, several feature vectors are created which can be merged in order to increase the security of the system and reduce vulnerability against brute force attacks.

REFERENCES

Ahmadyfard, A., & Modares, H. (2008). *Combining PSO and k-means to enhance data clustering.* Paper presented at the 2008 International Symposium on Telecommunications. 3(34), 3455-3543.

Akay, B., & Karaboga, D. (2009). *Parameter tuning for the artificial bee colony algorithm.* Paper presented at the International Conference on Computational Collective Intelligence. 6(4), 55-63.

Albuz, E., Kocalar, E., & Khokhar, A. A. (1998). Scalable image indexing and retrieval using wavelets. *Technical Report*, 11.

Alrifaee, M., Abdallah, M., & Al Okush, B. (2017). Al Okush. 2017. A Short Survey of IRIS Images Databases. *Int. J. Multimed. Its Appl, 9*(2), 01-14.

Alsmirat, M. A., Al-Alem, F., Al-Ayyoub, M., Jararweh, Y., & Gupta, B. (2018). Impact of digital fingerprint image quality on the fingerprint recognition accuracy. *Multimedia Tools and Applications*, 1-40.

Ansari, M. A., & Dixit, M. (2017). An Image Retrieval Framework: A Review. *International Journal of Advanced Research in Computer Science, 8*(5).

Anwar, A. (2016). An Iris detection and recognition system to measure the performance of E-security. BRAC University.

Arora, S., & Singh, S. (2013). *A conceptual comparison of firefly algorithm, bat algorithm and cuckoo search.* Paper presented at the 2013 International Conference on Control, Computing, Communication and Materials (ICCCCM).

Arthur, D., & Vassilvitskii, S. (2007). *k-means++: The advantages of careful seeding.* Paper presented at the Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms.

Asad, A. H., Azar, A. T., & Hassanien, A. E. (2017). A new heuristic function of ant colony system for retinal vessel segmentation. *International Journal of Rough Sets and Data Analysis (IJRSDA), 1*(2), 15-30.

Babich, A. (2012). Biometric Authentication. Types of biometric identifiers.

Bahmani, B., Moseley, B., Vattani, A., Kumar, R., & Vassilvitskii, S. (2012). Scalable k-means++. 5 (7): 622–633: March.

BATH Iris Database. (2017, October 2). *University of Bath*. Retrieved from http://www.smartsensors.co.uk/products/iris-database/

Barbu, T., & Luca, M. (2015). *Content-based iris indexing and retrieval model using spatial acces methods.* Paper presented at the 2015 International Symposium on Signals, Circuits and Systems (ISSCS).

Bastos-Filho, C. J., & Guimarães, A. C. (2015). Multi-objective fish school search. *International Journal of Swarm Intelligence Research (IJSIR), 6*(1), 23-40.

Bathla, G., Aggarwal, H., & Rani, R. (2018). A Novel Approach for Clustering Big Data based on MapReduce. *International Journal of Electrical & Computer Engineering (2088-8708), 8*(3).

Bay, H., Tuytelaars, T., & Van Gool, L. (2006). SURF: Speeded Up Robust Features. 2006. Dostupné na internete:*http://www.vision.ee.ethz.ch/~ surf/eccv06. pdf.*

Bechikh, S., Elarbi, M., & Said, L. B. (2017). Many-objective optimization using evolutionary algorithms: a survey *Recent Advances in Evolutionary Multi-objective Optimization* (pp. 105-137): Springer.

Bernard, F., Deuter, C. E., Gemmar, P., & Schachinger, H. (2013). Eyelid contour detection and tracking for startle research related eye-blink measurements from high-speed video records. *Comput Methods Programs Biomed, 112*(1), 22-37.

Biometrics Ideal Test. (2017 January 15). *Biometrics ideal test*. Retrieved from http://biometrics.idealtest.org/dbDetailForUser.do?id=4

Blasco, J., Chen, T. M., Tapiador, J., & Peris-Lopez, P. (2016). A survey of wearable biometric recognition systems. *ACM Computing Surveys (CSUR), 49*(3), 43.

Bose, A., & Mali, K. (2016). Fuzzy-based artificial bee colony optimization for gray image segmentation. *Signal, Image and Video Processing, 10*(6), 1089-1096.

Bouhmala, N., Viken, A., & Lønnum, J. (2015). Enhanced Genetic Algorithm with K-Means for the Clustering Problem. *International Journal of Modeling and Optimization, 5*(2), 150.

Bouras, C., & Tsogkas, V. (2010). *Assigning web news to clusters.* Paper presented at the 2010 Fifth International Conference on Internet and Web Applications and Services.

Bowyer, K. W., Hollingsworth, K., & Flynn, P. J. (2008). Image understanding for iris biometrics: A survey. *Computer vision and image understanding, 110*(2), 281-307.

Bsoul, Q., Al-Shamari, E., Mohd, M., & Atwan, J. (2014). *Distance Measures and Stemming Impact on Arabic Document Clustering.* Paper presented at the Asia Information Retrieval Symposium.

Bsoul, Q. W., & Mohd, M. (2011). Effect of ISRI stemming on similarity measure for Arabic document clustering. *In Asia Information Retrieval Symposium* . 584-593.

Burks, S., Harrell, G., & Wang, J. (2015). *On initial effects of the K-means clustering.* Paper presented at the Proceedings of the International Conference on Scientific Computing (CSC).

Cai, W., Chen, S., & Zhang, D. (2007). Fast and robust fuzzy c-means clustering algorithms incorporating local information for image segmentation. *Pattern recognition, 40*(3), 825-838.

Celebi, M. E. (2011). Improving the performance of k-means for color quantization. *Image and Vision Computing*, *29*(4), 260-271.

Celebi, M. E., Kingravi, H. A., & Vela, P. A. (2013). A comparative study of efficient initialization methods for the k-means clustering algorithm. *Expert systems with applications*, *40*(1), 200-210.

Chadha, A., & Kumar, S. (2014). *An improved K-means clustering algorithm: a step forward for removal of dependency on K.* Paper presented at the 2014 International Conference on Reliability Optimization and Information Technology (ICROIT).

Chaturvedi, D. (2008). Applications of genetic algorithms to load forecasting problem. *Soft Computing: Techniques and its Applications in Electrical Engineering*, 383-402.

Chaudhari, R. D., Pawar, A. A., & Deore, R. S. (2013). The historical development of biometric authentication techniques: A recent overview. *International Journal of Engineering Research & Technology (IJERT)*, *2*, 3921-3928.

Chen, D., Wan, S., Xiang, J., & Bao, F. S. (2017). A high-performance seizure detection algorithm based on Discrete Wavelet Transform (DWT) and EEG. *PloS one, 12*(3), e0173138.

Chen, M., Zhang, Y., & Lu, C. (2017). Efficient architecture of variable size HEVC 2D-DCT for FPGA platforms. *AEU-International Journal of Electronics and Communications, 73*, 1-8.

Cheng, H.-Y., & Yu, C.-C. (2014). Block-based cloud classification with statistical features and distribution of local texture features. *Atmospheric Measurement Techniques, 8*(3), 1173-1182.

Christmas, J., Keedwell, E., Frayling, T. M., & Perry, J. R. (2011). Ant colony optimisation to identify genetic variant association with type 2 diabetes. *Information Sciences, 181*(9), 1609-1622.

Cisty, M. (2010). Application of the harmony search optimization in irrigation. In *Recent Advances in Harmony Search Algorithm* 123-134

Claramunt, C., Schneider, M., Wong, R. C.-W., Xiong, L., Loh, W.-K., Shahabi, C., & Li, K.-J. (2015). Advances in Spatial and Temporal Databases: Presented at 14th International Symposium, , Hong Kong, China, August 26-28, 2015. Proceedings (9239): Springer.

Connolly, J.-F., Granger, E., & Sabourin, R. (2012). An adaptive classification system for video-based face recognition. *Information Sciences, 192*, 50-70.

Database of Indian Institute of Technology Kanpur. (2016 October 2). *Indian Institute of Technology Kanpur*. Retrieved from http://www.cse.iitk.ac.in/users/biometrics

Data, G. O., Han, I., & Kamber, M. (2010). Data mining: Concepts and techniques. *Morgan Kaufinann*.

Dalal, N., & Triggs, B. (2005). *Histograms of oriented gradients for human detection.* Paper presented at the international Conference on computer vision & Pattern Recognition (CVPR'05).

Daugman, J. (2006). Probing the uniqueness and randomness of IrisCodes: Results from 200 billion iris pair comparisons. *Proceedings of the IEEE, 94*(11), 1927-1935.

Delévacq, A., Delisle, P., Gravel, M., & Krajecki, M. (2013). Parallel ant colony optimization on graphics processing units. *Journal of Parallel and Distributed Computing, 73*(1), 52-61.

Dey, S., & Samanta, D. (2012). Iris data indexing method using Gabor energy features. *IEEE Transactions on Information Forensics and Security, 7*(4), 1192-1203.

Dey, S., & Samanta, D. (2014). *Unimodal and Multimodal Biometric Data Indexing*: Walter de Gruyter GmbH & Co KG.

Dey Sarkar, S., Goswami, S., Agarwal, A., & Aktar, J. (2014). A novel feature selection technique for text classification using naive bayes. *International scholarly research notices, 2014*.

Easwaramoorthy, S., Sophia, F., & Prathik, A. (2016). *Biometric Authentication using finger nails.* Paper presented at the 2016 International Conference on Emerging Trends in Engineering, Technology and Science (ICETETS).

Emad, T. K, & Norrozila, S. (2015). A New Biometric Template Protection Based On Secure Data Hiding Approach.

Emad, T. K, & Norrozila, S. (2015). Multibiometric systems and template security survey. *Journal of Scientific Research and Development, 2*(14), 38-46.

Eskandar, H., Sadollah, A., Bahreininejad, A., & Hamdi, M. (2012). Water cycle algorithm–A novel metaheuristic optimization method for solving constrained engineering optimization problems. *Computers & Structures, 110*, 151-166.

Falkenauer, E. (1998). *Genetic algorithms and grouping problems*: John Wiley & Sons, Inc.

Fan, J., Han, M., & Wang, J. (2009). Single point iterative weighted fuzzy C-means clustering algorithm for remote sensing image segmentation. *Pattern recognition, 42*(11), 2527-2540.

Farisi, O. I. R., Setiyono, B., & Danandjojo, R. I. (2016). A Hybrid Firefly Algorithm â€“Ant Colony Optimization for Traveling Salesman Problem. *Jurnal Buana Informatika, 7*(1).

Farnstrom, F., & Lewis, J. (2008). Fast, single-pass K-means algorithms.

Fierrez, J., Morales, A., Vera-Rodriguez, R., & Camacho, D. (2018). Multiple classifiers in biometrics. Part 2: Trends and challenges. *Information Fusion, 44*, 103-112.

Forgey, E. (1965). Cluster analysis of multivariate data: Efficiency vs. interpretability of classification. *Biometrics, 21*(3), 768-769.

Forsati, R., Mahdavi, M., Shamsfard, M., & Meybodi, M. R. (2013). Efficient stochastic algorithms for document clustering. *Information Sciences, 220*, 269-291.

Forster, E., Wallas, G., & Gide, A. (2017 April 7). *Cluster Analysis: see it 1st*. *Data Visualization*. Retrieved from https://apandre.wordpress.com/visible-data/cluster-analysis/

Fouad, M. (2012). *Towards Template Security for Iris-Based Biometric Systems.* Université d'Ottawa/University of Ottawa.

Fox, B., Xiang, W., & Lee, H. P. (2007). Industrial applications of the ant colony optimization algorithm. *The International Journal of Advanced Manufacturing Technology, 31*(7-8), 805-814.

Friedman, M., Last, M., Makover, Y., & Kandel, A. (2007). Anomaly detection in web documents using crisp and fuzzy-based cosine clustering methodology. *Information Sciences, 177*(2), 467-475.

Fun Ye* & Ching-Yi Chen. (2005). Alternative KPSO-clustering algorithm. *淡江理工學刊, 8*(2), 165-174.

Gadde, R. B., Adjeroh, D., & Ross, A. (2010). *Indexing iris images using the burrows-wheeler transform*. Paper presented at the 2010 IEEE International Workshop on Information Forensics and Security.

Gohberg, I., & Kreĭn, M. G. e. (1969). *Introduction to the theory of linear nonselfadjoint operators* (Vol. 18): American Mathematical Soc.

Ganorkar. S. & Rahman, M. (2013). Iris Recognition based on Neural Networks, *International Journal of Scientific & Engineering Research*. 4(12), 847-849.

Gragnaniello, D., Sansone, C., & Verdoliva, L. (2015). Iris liveness detection for mobile devices based on local descriptors. *Pattern Recognition Letters, 57*, 81-87.

Griffin, G., Holub, A., & Perona, P. (2007). Caltech-256 object category dataset.

Guo, Y., Li, W., Mileham, A. R., & Owen, G. W. (2009). Applications of particle swarm optimisation in integrated process planning and scheduling. *Robotics and Computer-Integrated Manufacturing, 25*(2), 280-288.

Gupta, D., & Choubey, S. (2015). Discrete wavelet transform for image processing. *International Journal of Emerging Technology and Advanced Engineering, 4*(3), 598-602.

Hamd, M. H., & Ahmed, S. K. (2018). Biometric system design for iris recognition using intelligent algorithms. *International Journal of Modern Education and Computer Science, 10*(3), 9.

Hanaa, A., S, A., & A.Farag, F. (2015). *Efficient enhancement and matching for iris recognition using SURF*. Paper presented at the 2015 5th national symposium on information technology: Towards new smart world (NSITNSW).

Huang, C.-L., Huang, W.-C., Chang, H.-Y., Yeh, Y.-C., & Tsai, C.-Y. (2013). Hybridization strategies for continuous ant colony optimization and particle swarm optimization applied to data clustering. *Applied Soft Computing, 13*(9), 3864-3872.

Iris Challenge Evaluation (ICE). (2017 April 2). *NIST*. Retrieved from https://www.nist.gov/programs-projects/iris-challenge-evaluation-ice

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters, 31*(8), 651-666.

Jain, A. K., Flynn, P., & Ross, A. A. (2008). *Handbook of biometrics*: Springer Science & Business Media.

Jiang, X. (2009). Fingerprint classification, *Encyclopedia of biometrics*. In S.Z. Li, & A.K. Jain (Eds.).439 – 445.

Jayaraman, U., Prakash, S., & Gupta, P. (2012). An efficient color and texture based iris image retrieval technique. *Expert systems with applications, 39*(5), 4915-4926.

Jia, Y., Wang, J., Zeng, G., Zha, H., & Hua, X.-S. (2010). Optimizing kd-trees for scalable visual descriptor indexing.

Jo, T. (2009). *Clustering news groups using inverted index based NTSO*. Paper presented at the 2009 First International Conference on Networked Digital Technologies.

Kakade, P., & Keche, I. (2017). Review on Content Based Image Retrieval (CBIR) Technique. *International Journal of Engineering and Computer Science*, 6(4). 20414-20416

Kao, Y.-T., Zahara, E., & Kao, I.-W. (2008). A hybridized approach to data clustering. *Expert systems with applications, 34*(3), 1754-1762.

Karaboga, D., & Basturk, B. (2008). On the performance of artificial bee colony (ABC) algorithm. *Applied Soft Computing, 8*(1), 687-697.

Kaur, H., & Pathania, S. (2016). Image enhancement and iris recognition using SIFT feature extraction. *Int. J. Adv. Res. Electron. Commun. Eng.(IJARECE), 5*(5), 1254-1256.

Kavati, I., Prasad, M. V., & Bhagvati, C. (2015). *Palmprint retrieval based on match scores and decision-level fusion.* Paper presented at the 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI).

Kavati, I., Prasad, M. V., & Bhagvati, C. (2017). Efficient Biometric Indexing and Retrieval Techniques for Large-Scale Systems: Springer.

Kavati, I., Prasad, M. V., & Bhagvati, C. (2016). Search space reduction in biometric databases: a review *Computer Vision: Concepts, Methodologies, Tools, and Applications* (pp. 1600-1626): IGI Global.

Kekre, H., Sarode, T. K., & Ugale, M. S. (2011). *An efficient image classifier using discrete cosine transform.* Paper presented at the Proceedings of the International Conference & Workshop on Emerging Trends in Technology.

Kerr, G., Ruskin, H. J., Crane, M., & Doolan, P. (2008). Techniques for clustering gene expression data. *Computers in biology and medicine, 38*(3), 283-293.

Khalaf, E. T., Mohammad, M. N., & Moorthy, K. (2018). Robust partitioning and indexing for iris biometric database based on local features. *IET Biometrics, 7*(6), 589-597.

Khalaf, E. T., Mohammad, M. N., Moorthy, K., & Khalaf, A. T. (2018). Efficient Classifying and Indexing for Large Iris Database Based on Enhanced Clustering Method. *Studies in Informatics and Control*, *27*(2), 191-202.

Khayam, S. A. (2003). The discrete cosine transform (DCT): theory and application. *Michigan State University, 114*.

Knitter-Piątkowska, A., & Guminiak, M. (2018). *Defect detection in plates using dynamic response signals and discrete wavelet transform.* Paper presented at the AIP Conference Proceedings.

Kumar, V., Chhabra, J. K., & Kumar, D. (2016). Automatic data clustering using parameter adaptive harmony search algorithm and its application to image segmentation. *Journal of Intelligent Systems*, *25*(4), 595-610.

Kuo, R., Syu, Y., Chen, Z.-Y., & Tien, F.-C. (2012). Integration of particle swarm optimization and genetic algorithm for dynamic clustering. *Information Sciences*, *195*, 124-140.

Lee, D., Park, S.-H., & Moon, S. (2013). Utility-based association rule mining: A marketing solution for cross-selling. *Expert systems with applications*, *40*(7), 2715-2725.

Lee, K. S., & Geem, Z. W. (2005). A new meta-heuristic algorithm for continuous engineering optimization: harmony search theory and practice. *Computer methods in applied mechanics and engineering, 194*(36-38), 3902-3933.

Leticia, C., Marcelo, E., Diego, I., Paolo, R. (2014). An efficient particle swarm optimization approach to cluster short texts. *Information Sciences*, *265*, 36-49.

Li, H., He, H., & Wen, Y. (2015). Dynamic particle swarm optimization and K-means clustering algorithm for image segmentation. *Optik, 126*(24), 4817-4822.

Luo, J., Liu, Q., Yang, Y., Li, X., Chen, M.-r., & Cao, W. (2017). An artificial bee colony algorithm for multi-objective optimisation. *Applied Soft Computing*, *50*, 235-251.

MacQueen, J. (1967). *Some methods for classification and analysis of multivariate observations.* Paper presented at the Proceedings of the fifth Berkeley symposium on mathematical statistics and probability.

Madhesiya, S., & Ahmed, S. (2013). Advanced technique of digital watermarking based on SVD-DWT-DCT and Arnold transform. *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), 2*(5), 1918-1923.

Mahdavi, M., & Abolhassani, H. (2009). Harmony K-means algorithm for document clustering. *Data Mining and Knowledge Discovery, 18*(3), 370-391.

Mahmud, M. S., Rahman, M. M., & Akhtar, M. N. (2012). *Improvement of K-means clustering algorithm with better initial centroids based on weighted average.* Paper presented at the 2012 7th International Conference on Electrical and Computer Engineering.

Mallat, S. (1989). Multifrequency channel decompositions of images and wavelet models. *IEEE Transactions On Acoustics, Speech, And Signal Processing, 37*(12), 2091-2110.

Manoj, V., & Elias, E. (2012). Artificial bee colony algorithm for the design of multiplier-less nonuniform filter bank transmultiplexer. *Information Sciences, 192*, 193-203.

Martíne, P., & Ramos, P. (2014). *A Feature Extraction Using SIFT with a Preprocessing by Adding CLAHE Algorithm to Enhance Image Histograms.* Paper presented at the 2014 International Conference on Mechatronics, Electronics and Automotive Engineering.

Mazumdar, J. B., & Nirmala, S. (2018). Retina Based Biometric Authentication System: A Review. *International Journal of Advanced Research in Computer Science, 9*(1).

Mehrotra, H. (2010). Iris identification using keypoint descriptors and geometric hashing. *Information Sciences, 12*, 13-23.

Mehrotra, H., & Majhi, B. (2013). Local feature based retrieval approach for iris biometrics. *Frontiers of Computer Science, 7*(5), 767-781.

Mehrotra, H., Majhi, B., & Gupta, P. (2010). Robust iris indexing scheme using geometric hashing of SIFT keypoints. *Journal of Network and Computer Applications, 33*(3), 300-313.

Mehrotra, H., Srinivas, B. G., Majhi, B., & Gupta, P. (2009). *Indexing iris biometric database using energy histogram of DCT subbands.* Paper presented at the International Conference on Contemporary Computing.

Meila, M., & Heckerman, D. (2013). An experimental comparison of several clustering and initialization methods. *arXiv preprint arXiv:1301.7401*.

Moghtadaiee, V., & Dempster, A. G. (2015). Determining the best vector distance measure for use in location fingerprinting. *Pervasive and Mobile Computing, 23*, 59-79.

Mohan, A. & Lindam, M. (2014). Image Enhancement Using DWT DCT and SVD. *International Journal of Engineering Research and Applications, 4*(4), 36-46.

Mohd, M., Bsoul, Q. W., Ali, N. M., Noah, S. A. M., Saad, S., Omar, N., & AZIZ, M. J. A. (2012). Optimal Initial Centroid in K-Means for Crime Topic. *Journal of Theoretical & Applied Information Technology, 45*(1).

Murthy, C. A., & Chowdhury, N. (1996). In search of optimal clusters using genetic algorithms. *Pattern Recognition Letters, 17*(8), 825-832.

Naik, A. (2017 April 7*). k-means clustering algorithm - Data Clustering Algorithms.*Retrieved from https://sites.google.com/site/dataclusteringalgorithms/k-means-clustering-algorithm.

Nair, S. A. H., & Aruna, P. (2015). Comparison of DCT, SVD and BFOA based multimodal biometric watermarking systems. *Alexandria Engineering Journal, 54*(4), 1161-1174.

Nayak, J., Naik, B., & Behera, H. (2016). A novel nature inspired firefly algorithm with higher order neural network: performance analysis. *Engineering Science and Technology, an International Journal, 19*(1), 197-211.

Nugroho, B. (2018). *Face Recognition of Robust Regression With Pre-processing Technique using CLAHE technique*. Paper presented at the Prosiding International conference on Information Technology and Business (ICITB).

Pan, J.-S., Snasel, V., Corchado, E. S., Abraham, A., & Wang, S.-L. (2014). Intelligent Data Analysis and Its Applications, Volume I: Proceeding of the First Euro-China Conference on Intelligent Data Analysis and Applications, June 13-15, 2014, Shenzhen, China (Vol. 297): Springer.

Parmar, P. A., & Degadwala, S. D. (2015). Fingerprint indexing approaches for biometric database: a review. *International Journal of Computer Applications, 130*(13).

Patel, V. (2018). Airport Passenger Processing Technology: A Biometric Airport Journey.

Patwal, P. S. (2012). A Content Based Indexing system For Image Retrieval. *Mobile Computing, 3*, 19-29.

Pedemonte, M., Nesmachnow, S., & Cancela, H. (2011). A survey on parallel ant colony optimization. *Applied Soft Computing, 11*(8), 5181-5197.

Pravin S., Kolhe S. R., Patil R. V. & Patil P. M. (2012). Performance Evaluation in Iris Recognition and CBIR System based on phase congruency. *International Journal of Computer Applications, 47*(14).

Puhan, N., & Sudha, N. (2008). *A novel iris database indexing method using the iris color*. Paper presented at the 2008 3rd IEEE Conference on Industrial Electronics and Applications.

Pyykkö, J. (2018). Online Personalization in Exploratory Search. *Engineering Journal, 10*(7), 61-74.

Radman, A., Jumari, K., & Zainal, N. (2012). Iris segmentation in visible wavelength environment. *Procedia Engineering, 41*, 743-748.

Rajaguru, H., & Prabhakar, S. K. (2017). KNN Classifier and K-Means Clustering for Robust Classification of Epilepsy from EEG Signals. *A Detailed Analysis. diplom. de*.

Rana, S., Jasola, S., & Kumar, R. (2011). A review on particle swarm optimization algorithms and their applications to data clustering. *Artificial Intelligence Review, 35*(3), 211-222.

Rathgeb, C., Breitinger, F., Baier, H., & Busch, C. (2015). *Towards bloom filter-based indexing of iris biometric data*. Paper presented at the 2015 international conference on biometrics (ICB).

Rathgeb, C., Breitinger, F., Busch, C., & Baier, H. (2013). On application of bloom filters to iris biometrics. *IET Biometrics, 3*(4), 207-218.

Rathgeb, C., & Uhl, A. (2010). *Iris-biometric hash generation for biometric database indexing*. Paper presented at the 2010 20th International Conference on Pattern Recognition.

Raykov, Y. P., Boukouvalas, A., Baig, F., & Little, M. A. (2016). What to do when k-means clustering fails: A simple yet principled alternative algorithm. *PloS one, 11*(9), e0162259.

Runkler, T. A. (2005). Ant colony optimization of clustering models. *International Journal of Intelligent Systems, 20*(12), 1233-1251.

Saad, I. A., & George, L. E. (2014). Robust and fast iris localization using contrast stretching and leading edge detection. *International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), 3*(2), 61-67.

Sadygov, R. G. (2014). Use of singular value decomposition analysis to differentiate phosphorylated precursors in strong cation exchange fractions. *Electrophoresis, 35*(24), 3498-3503.

Salcedo-Sanz, S., Pastor-Sánchez, A., Portilla-Figueras, J., & Prieto, L. (2015). Effective multi-objective optimization with the coral reefs optimization algorithm. *Engineering Optimization, 48*(6), 966-984.

Satapathy, S. C., Katari, V., Parimi, R., Malireddi, S., Misra, B., & Murthy, J. (2007). *A new approach of integrating PSO & improved GA for clustering with parallel and transitional technique*. Paper presented at the Third International Conference on Natural Computation (ICNC 2007).

Severo, E., Laroca, R., Bezerra, C. S., Zanlorensi, L. A., Weingaertner, D., Moreira, G., & Menotti, D. (2018). *A benchmark for iris location and a deep learning detector evaluation.* Paper presented at the 2018 International Joint Conference on Neural Networks (IJCNN).

Shunye, W. (2013). *An improved k-means clustering algorithm based on dissimilarity.* Paper presented at the Proceedings 2013 International Conference on Mechatronic Sciences, Electric Engineering and Computer (MEC).

Si, Y., Mei, J., & Gao, H. (2012). Novel approaches to improve robustness, accuracy and rapidity of iris recognition systems. *IEEE transactions on industrial informatics, 8*(1), 110-117.

Singh, A., Yadav, A., & Rana, A. (2013). K-means with Three different Distance Metrics. *International Journal of Computer Applications, 67*(10).

Singh, U. K., Prajapati, R., & Kumar, T. (2016). Geological stratigraphy and spatial distribution of microfractures over the Costa Rica convergent margin, Central

America–a wavelet-fractal analysis. *Geoscientific Instrumentation, Methods and Data Systems*, *7*(2), 179-187.

Song, Y., McLoughLin, I., & Dai, L. (2015). *Deep bottleneck feature for image classification.* Paper presented at the Proceedings of the 5th ACM on International Conference on Multimedia Retrieval.

Sridhar, B. (2017). A Blind Image Watermarking Technique Using Most Frequent Wavelet Coefficients. *International Journal on Smart Sensing & Intelligent Systems, 10*(4).

Stokkenes, M., Ramachandra, R., Sigaard, M. K., Raja, K., Gomez-Barrero, M., & Busch, C. (2016). *Multi-biometric template protection—A security analysis of binarized statistical features for bloom filters on smartphones.* Paper presented at the 2016 Sixth International Conference on Image Processing Theory, Tools and Applications (IPTA).

SmartSensores. (2014 July). *CASIA Iris Database*. Retrieved from www.smartsensors.co.uk/products/iris-database.

Svagerka, M. (2018). On the Complexity of Recognizing Similarities between Streams. ETH Zurich.

Swapna, C. S., Kumar, V. V., & Murthy, J. (2016). Improving Efficiency of K-Means Algorithm for Large Datasets. *International Journal of Rough Sets and Data Analysis (IJRSDA), 3*(2), 1-9.

Taha, K., & Norrozila, S. (2015). A Survey of Multi-Biometrics and Fusion Levels. *Indian Journal Of Science And Technology*, 8(32). 1-10.

The Bubble Sort Algorithm – Sorting One-Dimensional Arrays with Numeric Values. (2017 March 26). *Aristides S. Bouras*. Retrieved from http://www.bouraspage.com/repository/algorithmic-thinking/the-bubble-sort-algorithm-sorting-one-dimensional-arrays-with-numeric-values.

The Unique Identification Authority of India (UIDAI). (2017 April 15). *UIDAI*. Retrieved from https://uidai.gov.in/

Tidke, B., Mehta, R., & Rana, D. (2012). A novel approach for high dimensional data clustering. *International Journal of Engineering Science and Advanced Technology (IJESAT), 2*(3).

Tilahun, S. L. & Ong, H. C. (2012). Modified firefly algorithm, *Journal of Applied Mathematics*, 467631(12).

Tractica.com. (2017 February 15). *Iris Recognition Biometrics Market*. Retrieved from https://www.tractica.com/newsroom/press-releases/iris-recognition-biometrics-market-to-increase-to-4-1-billion-worldwide-by-2025/

Uludag, U., Pankanti, S., Prabhakar, S., & Jain, A. K. (2004). Biometric cryptosystems: issues and challenges. *Proceedings of the IEEE, 92*(6), 948-960.

University of Palackeho and Olomouc. (2015 April 2). *iris databases* . Retrieved from *http://phoenix.inf.upol.cz/iris.*

Velmurugan, T., & Santhanam, T. (2011). An experimental approach. *Information Technology Journal, 10*(3), 478-484.

Vielhauer, C., Dittmann, J., Drygajlo, A., Juul, N. C., & Fairhurst, M. (2011). *Biometrics and ID Management: COST 2101 European Workshop, BioID 2011, Brandenburg (Havel), March 8-10, 2011, Proceedings* (Vol. 6583): Springer Science & Business Media.

Wang, J. Z., Li, J., & Wiederhold, G. (2001). Simplicity: Semantics-sensitive integrated matching for picture libraries. *IEEE Transactions on Pattern Analysis & Machine Intelligence*(9), 947-963.

Wang, X., Qiu, S., Liu, K., & Tang, X. (2014). Web image re-ranking usingquery-specific semantic signatures. *IEEE transactions on pattern analysis and machine intelligence, 36*(4), 810-823.

Wei, X. (2010). *Improved ant colony algorithm based on information entropy.* Paper presented at the 2010 International Conference on Computational and Information Sciences.

Wolfson, H. J., & Rigoutsos, I. (1997). Geometric hashing: An overview. *IEEE computational science and engineering, 4*(4), 10-21.

Wu, D.-s., & Wu, L.-n. (2002). *Image retrieval based on subband energy histograms of reordered DCT coefficients.* Paper presented at the 6th International Conference on Signal Processing, 2002.

Xiaoming, S., Ning, Z., Haibin, W., Xiaoyang, Y., Xue, W., & Shuang, Y. (2018). Medical Image Retrieval Approach by Texture Features Fusion Based on Hausdorff Distance. *Mathematical Problems in Engineering, 2018.*

Yang, F., Sun, T., & Zhang, C. (2009). An efficient hybrid data clustering method based on K-harmonic means and Particle Swarm Optimization. *Expert systems with applications, 36*(6), 9847-9852.

Yang, X.-S. (2010). Nature-inspired metaheuristic algorithms: Luniver press.

Yu, H., Jia, M., Cheng, X., & Jiang, Q. (2013). *Optimized k-means clustering algorithm based on artificial fish swarm.* Paper presented at the Proceedings 2013 International Conference on Mechatronic Sciences, Electric Engineering and Computer (MEC).

Zgrzywa, A., Choroś, K. & Siemiński, A. (2017). *Multimedia and Network Information Systems*. Proceedings of the 10th International Conference MISSI 2016.

Zhao, M., Tang, H., Guo, J., & Sun, Y. (2014). Data clustering using particle swarm optimization *Future Information Technology* (pp. 607-612): Springer.

APPENDIX A
EXPERIMENTAL RESULTS

Table A.1 Performance rates for the system using four large created datasets

Database name	Database size	1D feature vector using K-means Algorithm		2D feature vector using K-means Algorithm		Indexing method using WKIFA Algorithm	
		BM %	PR %	BM %	PR %	BM %	PR %
DS1	5000 iris images	0.02891	69.648	0.033277	27.852	0.001592	33.984
		0.030332	24.492	0.112988	7.98	0.053067	14.4
		0.218009	11.208	0.256269	3.108	0.066599	8.088
		0.291706	8.664	0.361628	1.14	0.141423	3.444
		0.519194	4.608	0.374895	0.732	0.146332	2.34
		0.63981	2.736	0.410936	0.3	0.187326	1.212
		0.719668	1.176	0.467209	0.156	0.267855	0.144
DS2	10000 iris images	0.057089	110.8796	0.023109	44.34038	0.00176	54.10253
		0.059897	38.99126	0.078464	12.70416	0.058669	22.9248
		0.430508	17.84314	0.177965	4.947936	0.073629	12.8761
		0.576039	13.79309	0.251131	1.81488	0.156352	5.482848
		1.025265	7.335936	0.260345	1.165344	0.161779	3.72528
		1.263448	4.355712	0.285373	0.4776	0.2071	1.929504
		1.421145	1.872192	0.324452	0.248352	0.29613	0.229248
DS3	20000 iris images	0.070813	195.1481	0.009118	78.03907	0.001222	95.22045
		0.074296	68.62462	0.03096	22.35932	0.040743	40.34765
		0.534004	31.40393	0.07022	8.708367	0.051131	22.66194
		0.714521	24.27584	0.099089	3.194189	0.108578	9.649812
		1.271743	12.91125	0.102725	2.051005	0.112347	6.556493
		1.567186	7.666053	0.1126	0.840576	0.14382	3.395927
		1.762794	3.295058	0.12802	0.4371	0.205647	0.403476
DS4	30000 iris images	0.049907	238.471	0.002944	95.36374	0.000482	116.3594
		0.052362	83.85929	0.009997	27.32309	0.016076	49.30483
		0.376353	38.3756	0.022673	10.64162	0.020175	27.69289
		0.503577	29.66508	0.031995	3.903299	0.042842	11.79207
		0.896293	15.77755	0.033169	2.506328	0.044329	8.012034
		1.104514	9.367917	0.036357	1.027184	0.056747	4.149823
		1.242373	4.026561	0.041336	0.534136	0.081143	0.493048

APPENDIX B
RESEARCH PUBLICATION

1. **Emad Taha Khalaf**, Muamer N. Mohammad and Kohbalan Moorthy. Robust Partitioning and Indexing for Iris Biometric Database Based on Local Features. ***IET Biometrics***. 7(6) 589 – 597. 2018. (**ISI**: **Impact Factor: 1.936 /Q2**). http://ietdl.org/t/1HFhob.

2. **Emad Taha Khalaf**, Muamer N. Mohammad and Kohbalan Moorthy. Efficient Classifying and Indexing for Large Iris Database based on Enhanced Clustering Method. ***Studies in Informatics and Control*** . 27(2) 191-202, June 2018. (**ISI**: **Impact Factor: 1.020**). https://sic.ici.ro/wp-content/uploads/2018/07/Art.-7-Issue-2-2018-SIC.pdf.

3. **Emad Taha Khalaf** and Norrozila Sulaiman. A Survey of Multi-Biometrics and Fusion Levels. ***Indian Journal of Science and Technology***, Vol 8(32), DOI: 10.17485/ijst/2015/v8i32/91488, November 2015. **(SCOPUS)**. http://www.indjst.org/index.php/indjst/article/view/91488/68641.

4. **Emad Taha Khalaf,** Mohammed MN, Sulaiman N. Iris Template Protection based on Enhanced Hill Cipher. *InProceedings of the 2016 International Conference on Communication and Information Systems* 2016 Dec 16 (pp. 53-57). **(ACM *Digital Library*)**. https://dl.acm.org/citation.cfm?id=3023938**.**

5. **Emad Taha Khalaf** and Norrozila, S. A new biometric template protection based on secure data hiding approach. *ARPN Journal of Engineering and Applied Sciences*, 10(2). 2015. **(SCOPUS)**. http://www.arpnjournals.com/jeas/research_papers/rp_2015/jeas_0215_1480.pdf.

6. **Emad Taha Khalaf**, Muamer N. Mohammad, Kohbalan Moorthy and Raed Abdulkareem. Biometric Template Protection based on Hill Cipher Algorithm with Two Invertible keys. the 5th International Conference on Software Engineering & Computer Systems (ICSECS17), Langkawi, Malaysia, 22 - 24 November 2017. **(UMP / FSKKP)**. http://umpir.ump.edu.my/id/eprint/19976.

7. **Emad Taha Khalaf** and Norrozila Sulaiman. Multibiometric Systems and Template Security Survey , *Journal of Scientific Research and Development*. 2 (14): 38-46, ISSN 1115-7569, 2015. http://jsrad.org/wp-content/2015/Issue%2014,%202015/7jj.pdf.

8. **Emad Taha Khalaf** and Norrozila Sulaiman. A New Secure Storing System for Biometric Templates based Encryption and Concealment. *Journal of Applied*

Sciences, 69404-JAS-ANSI. 2015. https://scialert.net/abstract/?doi=jas.2015.773.782.

9. **Emad T Khalaf**, Norrozila Sulaiman and Muamer N. Mohammad, Anti-Forensic Steganography Method Based On Randomization, *Global Journal on Technology*, Vol.(04), pp. 447-453. 2013. http://www.academia.edu/35469705/Anti-Forensic_Steganography_Method_Based_On_Randomization.

10. **Norrozila** Sulaiman, Muamer N. Mohammed, **Emad T. khalaf**. Information Hiding for Electronic Holy Quran Image Protection based on Randomisation. Taibah University *International Conference on Advances in Information Technology for the Holy Quran and Its Sciences* (NOORIC1435/2013), Al-Madinah Al-Munawwarah, Saudi Arabia. 2013. http://fskkp.ump.edu.my/sysnets2/index.php/en/publication.

11.**Raed** Albukamrh, Muamer N. Mohammed, Mohammed Ariff Bin Ameedeen and Emad Taha Khalaf . Dynamic Load Balancing Model Based on Server Status (DLBS) for Green Computing. the 5th International Conference on Software Engineering & Computer Systems (ICSECS17), Langkawi, Malaysia, 22 - 24 November 2017. **(UMP / FSKKP)**. http://icsecs.ump.edu.my/index.php/en/program/program-book/file.