INTEGRITY VERIFICATION FOR TEXT CONTENT IN E-LEARNING SYSTEMES

A
THESIS
SUBMITTED TO

GALGOTIAS UNIVERSITY
GREATER NOIDA

IN FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY
IN
COMPUTER APPLICATIONS

BY

MR. FATEK SAEED KHALID SAEED

Regd. No. – 1413302002

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING
GALGOTIAS UNIVERSITY, GREATER NOIDA, U.P., INDIA

DECEMBER, 2019

CANDIDATE'S DECLARATION

I hereby certify that the work which is being presented in the thesis, entitled "**Integrity Verification for Text Content In E-Learning Systems**" in fulfillment of the requirements for the award of the degree of Doctor of Philosophy (Ph.D.) in Computer Applications under the School of Computer Science and Engineering**,** and submitted in Galgotias University, Greater Noida is an authentic record of my own work carried out during a period from March, 2015 to December, 2019 under the supervision of Prof.(Dr.) Anurag Dixit.

The matter embodied in this thesis has not been submitted by me for the award of any other degree of this or any other University/Institute.

(Fatek Saeed Khalid Saeed)

This is to certify that the above statement made by the candidate is correct to the best of our knowledge.

Prof. (Dr.) Anurag Dixit
Supervisor
School of Computer Sciences & Engineering

The Ph.D. Viva-Voice examination of **Mr. Fatek Saeed Khalid Saeed** Research Scholar, has been held on _____ .

Supervisor External Examiner

ABSTRACT

The use of E-Learning and online educations have been growing up due to its boundless benefits to every individuals. It has evolved from a knowledge transfer model to a great intellect, speedy and interactive proposition capable of advanced decision-making abilities. The e-learning methods have so many applications like e-banking, e-commerce and e-government. In e-learning methods the learning contents may affect by integrity attacks, security threats and tampering attacks. To avoid these kinds of attacks and provide security for the learning contents, some of the watermarking techniques are studied in this thesis. The integrity verification and tampering detection are providing security for the original contents.

To assure the integrity of the content, the original document is enforced to detection and verification attempt for preventing from tampering. The documents are usually watermarked with an efficient watermarking technique. The tampering in document alters or destructs the watermark and can misuse the document by the unauthorised user. Initially this thesis studied about the integrity verification and watermarking techniques for e-learning system. The watermarking methods have several classifications such as text watermarking, image watermarking, audio watermarking, video watermarking, perceptivity watermarking, and robust watermarking. But zero watermarking provides most accurate results for the tampering detection and integrity verification.

The watermarking technique has some stages including watermark embedding, watermark extraction, watermark detection and possible attacks during the watermarking process. The watermark embedding process done by the embedding algorithm and the watermark extraction process extracts the watermark from the text by extraction algorithm. The biometric system is briefly described and analysed in this thesis. In biometric cryptosystem there are two phases, Enrollment phase and authentication phase. The enrollment phase processed the encoding and encryption and the authentication phase of the biometric cryptosystem performs the decoding and matching process.

The biometric cryptosystem is utilized for the template protection. In this method error correcting output code matrix is adopted for the encoding and decoding process. The biometric cryptosystem has two processes including key-binding and key generation. The presented biometric cryptosystem in this thesis performs the template protection by the

help of error correcting output codes scheme. The discriminant binarization transformation, chaos feature permutation and key generation are the procedure for the template protection method. Markov model combined with zero watermarking of English text documents for integrity verification are presented in this thesis.

The word level and letter level based Markov model with zero watermarking technique is analysed. 3 gram LNMZW developed and 4, 5 gram WNMZW algorithms and described. In these algorithms watermark embedding process is logical because original text file is not customized. In this thesis WNMZW algorithm based watermarking process is presented. The Markov chain is utilized to analyse the contents of the English text documents. The third order of letter mechanism Markov model and fifth order of word mechanism Markov model are used to produce the watermark key which is logically embedded within the plain text file or had backup in the watermark database. After that the generated watermark key utilized to match the watermark generated from the attacked document for detecting the tampering in the document and authenticating its content.

Based on the Ngram orders of Markov model in word mechanism each word of the text is considered as 1 gram or one order of Markov model. The each word of the text represented as separate state in Markov chain. Based on this procedure the 5 gram of Markov model is improved and compared its performance with other algorithms with different attacks in thesis. Initially the Markov chain is generated. Later, the watermark generation and original watermark detection process are done. Based on the text states and transitions state the watermark is generated. The presented LNMZW3 and WNMZW5 approach compared with different dataset sizes and the performance of the presented approach are obtained with different attacks.

In this thesis, a zero watermarking technique is proposed based on the hybrid structural component and word length (HSW). This HSW based method has watermark embedding process and extraction process. The plain text is obtained from the data owner and then the watermark is generated. This watermark generation process is on the basis of using the characteristics of the text document. Based on the watermarking approaches, the content of the digital medium is customized with the general watermarking approaches. The properties of the original data did not modified in this approach but it utilizes the original data in order to generate the watermark information. The text and watermark are registered to the certifying authority along with the original plain text.

Later, it utilized for the procedure of pattern matching which detects the tampering in text document. In order to form the groups based on group size, the text partition is performed. Moreover, to detect the maximum occurring list the word or letter from each partition is checked. For the designing of watermark key on the basis of watermark, the occurring list is used. The proposed HSW based watermarking method generates the watermark on the basis of the characteristics of the plain text instead of embedding the text itself. After watermarking the data whether it attacked or not, the extraction process is executed.

The watermark pattern generated by the watermark, which is detected from the original text with the use of an extraction algorithm of HSW. After the extraction the each detected pattern compared with the registered pattern with the certifying authority for the detection of tampering attacks. The tampered document detected by the pattern matching procedure. By these proposed HSW based watermarking approach the attacks including insertion, deletion and reorder attacks are avoided.

Dedication.....

I dedicate this research work,

To my Mother, my Wife, my Son Mohammad, and to my Daughters: Ansam, Sara, Sham.

ACKNOWLEDGEMENT

In the name of Allah, the Most Merciful, the Most Graceful.

The research work of this thesis is the outcome of consistent association with many colleagues and people. Therefore, I would like to take this opportunity to express my appreciation and sincere thanks to those who were always ready to extend their help to make this PhD thesis possible.

First and foremost, thanks and praises to the Almighty, Allah, for his blessings throughout my research work and for providing me with all the strength, happiness and good health to carry out this study.

Very special thanks to the Shri Sunil Galgotia Chancellor and Mr. Dhruv CEO, Galgotias University for giving me the opportunity to carry out my doctoral research and for their support.

I owe my deepest gratitude towards Prof.(Dr.) Preeti Bajaj Vice Chancellor , Galgotias University for providing favorable academic environment to carry out my research.

I feel privileged to have worked under the guidance of Prof(Dr.) Anurag Dixit, School of Computer Science and Engineering, Galgotias University. I express my sincere gratitude for spent much of his time in correcting my work thoroughly, his precious instructions and comments that help me in completing my PhD dissertation, His guidance enabled me to complete my work and make it what it is.

From the core of my heart, I would like to thank External member of Doctoral Committee Prof.(Dr.) D.K. Lobiyal, Ex-Dean School of Computer and Systems Sciences, Jawaharlal Nehru University, New Delhi, for his instructions and insightful comments that helped me to improve my work.

I am thankful and grateful to the Prof.(Dr.)S. Raju , Dean School of Computer Sciences & Engineering, for his supports and motivations. It's from my heart to thank Prof.(Dr.) Naresh Kumar, Research Coordinator-SCSE. I am also thankful to all the

faculty members in the School of Computer Sciences & Engineering at Galgotias University.

I express my gratitude to Dr. Tofik Abdulhamed, Professor of Surgery, and Manager of Modern surgical centre, Yemen, who encouraged me to finish my Ph.D. degree and who was kind enough to look after my family in Sanaa, while being in India for the last five years.

It gives me pleasure to thank Prof. Adnan Zain, Prof. Ghalib Algafari, Dr. Fahd Alwesabi, Dr. Ahmed Alhusami, Dr. Faiz Aljobai, Dr. Ali Kosheh, MD Munaser who kindly advised and supported me to develop my career and pursue my Ph.D. degree.

I must not forget to thank my family: My mother whose prayers guide me always in my entire life. Special thanks to my loving wife for her patience which can't be expressed in words. My heartfelt thanks are also due to beloved sons, daughters, brothers, sister for their love, affection and care. My uncles Mr. Saif Ahmed, cousins Mr. Sami Abdullah , Mr. Abdul Majeed Khalid and other relatives, who kept praying to Allah to help me achieving my dream all the time. Last but not the least, blessings and prayers are also to the soul of my Father, and to my uncle Mr. Mohammad Ahmed for all his supports.

It gives me pleasure to thank the ambassador and the Consul of Yemen embassy at Delhi, and I have to mention Mr. Mohammed Al Ameri Yemenis representative at India for all his support, moreover to the Cultural attach, for their nice help in supporting me to accomplish this work, together with the Yemeni Diplomatic mission in the Yemeni embassy as a whole.

Finally, I want to offer my gratitude to the officials at Ministry of Higher Education and Scientific Research for granting me the financial support to carry out this study.

Thanks Allah for everything in my life.

FATEK SAEED KHALID SAEED

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Abbreviation	Titles
AFP	Average Frequency Proposition
ARM	Attack via Record Multiplicity
ASCII	American Standard Code for Information Interchange
BCS	Biometric Crypto-System
CA	Certification Authority
CGA	Compatible Genetic Algorithm
DBT	Discriminant Binarization Transformation
DCT	Discrete Cosine Transform
DFT	Discrete Fourier Transform
DWM	Digital Watermark
DWT	Discrete Wavelet Transform
ECOC	Error Correcting Output Code
ELSS	E-learning System Success
EW	Extracted Watermark
GS	Group Size
HAS	Human Auditory System
HSW	Hybrid Structural component and Word Length
IDWT	Inverse Discrete Wavelet Transform
LCMS	Learning Content Management System

LDS	Learning Design System
LNMZW	Letter level Ngram Markov model Zero Watermarking
LMS	Learning Management System
LPM	Log Polar Mapping
LSS	Learning Support System
LST	Large Size Text
MST	Middle Size Text
NS	Next State
OW	Original Watermark
PDF	Portable Document Form
PM	Pattern Matching
Pr	Partition Size
PS	Present State
PSO	Particle Swarm Optimization
RSA	Rivest-Shamir-Adleman Algorithm
RST	Rotation Scaling and Rotation invariants
SST	Small Size Text
SVD	Singular Value Decomposition
SVM	Support Vector Machine
TAM	Technology Acceptance Model

TP	Total Pattern
TPB	Theory of Planned Behaviour
TUM	Technology Use Model
WA	Watermark Alphabet
WDR	Watermark Distortion Rate
WELS	Web based E-learning System
WNMZW	Word level Ngram Markov model Zero Watermarking
WMP	Watermark Pattern
MOV	Maximum Occurring Vowels
MONV	Maximum Occurring NON-Vowels
MOFL	Maximum Occurring File List
MDC	Maximum Detection Count

LIST OF SYMBOLS

Chapter 3

$f(x,y)$	Image in space domain
$F(u,v)$	Frequency domain using DFT
$W(u,v)$	Watermark used in DFT domain
$\hat{F}(u,v)$	Watermarked image in DFT domain
$\hat{F}_i(u,v)$	Segment of original image in DFT domain
$\hat{W}(u,v)$	Recovered suspected watermark
$N \times M$	Image size
$M_{N \times n}$	ECOC matrix
N_C	Number of the classes
M	Encoding matrix
E	Extension matrix
M_t	Encoding matrix in iteration t
E_t	Extension matrix in iteration t
k	Number of extended columns
F	Discriminant function
x	Training template
W_j	Weight vector

a_j	Threshold
W_q	Binary template
n	Length
key_q	User key
P_q	Permuted binary template

Chapter 4

T_1	Original text document
$T(A)$	Attached document
$PM(p)$	No of matched primary patterns
$PM(s)$	No of matched secondary patterns
U_s	Unique states
T_s	Possible transitions for each state
P_t	Possible transitions for all states
P_s	Possible sets
n	Total count of all character set
R_s	Count of all repeated states within Markov chain matrix
$M[i, j]$	Markov transition matrix
i	Present state
j	Next state

WMP_O	Original watermark patterns
WMP_A	Attacked watermark patterns
$Ngram$	Order level of Markov model
S_w	Weight of the states correctly matched

Chapter 5

T_e	Text content
G_s	Group size
N_g	Number of groups
L_r	Frequency list

PREFACE

Plain text digital watermarking has a most essential area in scientific research, which may open the new skyline in digital content security, where recently consider a backbone and best solution to address content authentication problems and protection it.

This thesis aims to design and concentrate on the watermarking techniques to improve the integrity verification and tampering detection of learning contents on e-learning systems. The organization of the thesis describes as follows.

In chapter 1, introduction to e-learning, e-learning systems, barriers and advantages of e-learning are provided. E-learning security threats and issues also provided in this chapter.

In chapter 2, literature survey for the e-learning systems with different watermarking approaches for integrity verification is provided.

In chapter 3, text watermarking and integrity verification is mainly concentrated. The biometric cryptosystem with error correcting output code matrix is deals in this chapter.

In chapter 4, combined Markov model and zero watermarking technique for integrity verification of PDF English text documents is described and this method compared with existing methods.

In chapter 5, the proposed HSW based zero watermarking technique is presented and compared with existing method.

In chapter 6, the general summary and major conclusions of all chapters are provided with conclusion and future scope.

LIST OF PUBLICATIONS

The author has a number of publications regarding his research work as listed below:

1. F. Saeed and A. Dixit, *"Integrity Verification & Temper Detection of English Documents using Hybrid Structural Component and Word Length, IJEAT ISSN: 2249 – 8958, Volume-9 Issue-1, pp. 7073–7078, October 2019.*

2. Fatek Saeed and Prof.(Dr). Anurag Dixit, *"Literature Survey of Digital Watermarking Techniques for Content integrity",* in Proceeding of the International Conference on Advanced Scientific Innovation in Science, Engineering and Technology (ICASISET 2019), Bharath Institute of Higher Education and Research, Chennai, India, 19-23 April 2019 pp 1338-1343 .

3. Fatek Saeed and Prof.(Dr). Anurag Dixit, "*COMBINED MARKOV MODEL AND ZERO WATERMARKING FOR INTEGRITY VERIFICATION OF ENGLISH TEXT DOCUMENTS"*, Advances in Intelligent Systems and Computing, ISSN 2194-5357, Vol. 989, Springer Nature Singapore 2020 pp857-864.

4. Fatek Saeed and Prof.(Dr). Anurag Dixit, "*HYBRID HSW BASED ZERO WATERMARKING FOR TAMPERING DETECTION OF TEXT CONTENTS"*, Lecture Notes on Data Engineering and Communications Technologies, ISSN 2367-4512, Vol 31, Springer Cham Switzerland AG 2020pp 820-826.

5. Fatek Saeed and Anurag Dixit, "*Digital Watermarking Techniques for Content integrity on e-learning Systems*", in Proceeding of the 2018 IEEE International Conference on Computing, Power and Communication Technologies(GUCON), Galgotias University, G Noida, India, pp 757- 762.

6. Fatek Saeed and Prof. Anurag Dixit, " *A Decision Support System Approach for Accreditation & Quality Assurance Council at Higher Education Institutions in Yemen*, in Proceeding of IEEE 3rd International Conference on MOOCs, Innovation and Technology in Education (MITE) October 2015, pp 163-168.

CHAPTER 1
E-LEARNING INTRODUCTION

CHAPTER 1

E-LEARNING INTRODUCTION

1.1 Introduction:

The growth in internet technology made a big impact in the society. People around the world can contact through the internet region. The education field became smarter with the implementation of internet. The web-based learning or online-learning determined as e-learning. The term e-learning utilized to refer the other technologies and web based uses in terms of teaching and learning process enhancing. The e-learning have same characteristics and functionalities of another e-services including e-banking, e-commerce, e-library, emails and e-government. The e-services behaviours vary with respect to the roles and requirements of the users. The users of e-learning focused on to gain the advantages of e-learning process concerning teaching and learning purposes. In e-learning technique the user may spend more time when compared to other e-services. The e-learning is interactive based learning in which the subject content were available on-line and gives automated feedback to the user activities towards learning. However, e-learning has become viral in high level education institutions [1].

In the era of digital world, the usage of personal computers and internet communication has trend, hence the e-learning have increased dramatically in the recent years. The concept of e-learning varies from the existing educational system. The learner-oriented environment provided for the teacher and students and also it delivers digital materials [2]. E-learning provides a condition to recognize the life-long learning principles to construct a real time learning community. To present the knowledge, communicate with it and shares it with others is required tools and techniques in the process of learning. E-learning is the very essential tool used to support the learning system to accomplish its goal. These types of technologies are helped to produce and distribute individualized, comprehensive, dynamic learning content that promotes learning anytime and anywhere. Based on that, students are

capable to present in online based learning by PCs, mobile phones and other handheld electronic devices.

The term e-Learning is introduced during the 1980's, at the same time the online learning was also delivered. Many authors defines the e-learning especially, some others apply a particular definition or view of e-learning. In 21st century the e-learning systems are made big impact on the educational field. The education system does not changed always by the technology. Education is one of the functions of the social process in which the communication of the system plays major roles. Therefore, the learning and gaining knowledge is important part for any society, organization or person.

The structural infrastructure is a part from e-learning systems which leverages knowledge diffusion and achievement. The system of e-learning also permits the socialization within a knowledge sharing context [3]. The e-learning have several definitions in the literature, based on the content, communication and technology. The general description recommended that the e-learning covers a wide range of implementations and operations including computer based learning, web-based learning, smart classrooms and digital collaboration. The knowledge or information can be acquired by the implementation of communication technology supported by e-learning.

The way of teaching and learning methods were changed completely by the technology development. In education field the electronic learning (e-learning) is the advanced approach, highlights learner-oriented and lifelong teaching-learning operations. Generally, the e-learning refers to utilize the network technology to deliver instruction and information to individuals, initially through the internet or over the intranet. The requirements for learning is technologically advanced region, achieved by the characteristics of e-learning. Based on the business demand of e-learning the growth is highly increased recently.

For promoting the teaching and learning operations, e-learning systems have popular and prominent tools that allow convenient learning. The e-learning system can be simply described as an information system which can effectively integrate a wide range of different instructional materials such as text, audio and video mediums. The instructions conveyed through e-mail, live chat sessions, online conversations and assignments. Moreover, the e-learning systems allow industrial information delivery among trainers and learners through the internet, intranet, extranet, CD-ROMs and satellite broadcasts interactive TV [4].

The communication among the devices in an e-learning system can be classified into two types: asynchronous and synchronous on the basis of temporal restrictions established on learning process. The learners log on to an e-learning system for the purpose of downloading materials or communicate between instructors and peers at any time in the asynchronous e-learning. Generally, this e-learning simplified based on the media including e-mail and discussion boards to help collaboration works among instructors and learners even though they may not be in online at the same time.

The main strategic perception of the e-learning is flexibility. Indeed, several users will use the online courses because of its asynchronous characteristics. The synchronous characteristics permit users to have online sessions for real-time communication between the instructor and learner. Generally, this e-learning helps in different kind of media based applications including video conferencing and chat, and also permits users to upgrade communities of learners. Finally, the benefits from the e-learning are independent of geographic location and the benefits from traditional education such as face to face interaction are combined [5].

The e-learning system also implemented in real time applications. Recently, e-learning provides opportunity to everyone to become a learner. The main problem associated with traditional learning is distance, but e-learning provides lectures anytime and anywhere. This concept of e-learning promotes the process of learning and avoids the problems in learning. In selection of online courses, the adaptabilities by which e-

learning can provide to the learners are the major inspiring factor. Furthermore, the technology utilized in learning will offer several benefits including quality improving of learning, education and training access improving, decrease the costs correlated with education and cost-efficient improvement of education.

The system of e-learning offers a stages of a well-designed, student-centred, attractive, communicative, simply accessible, convenient and deliberately delivered and also simplified e-learning environment. Furthermore, the students can avoid the cost and time spent for getting right materials needed for their study and also they can avoid the costs for the materials to be printed, due to online reading facility of learning materials. Moreover, the access to learning materials increases in e-learning. It also activates students to have larger access of limited resources including e-books and e-journals. This supports the learners to upgrade their learning method. The students can take charge of their own life-long learning based on avoiding the issues of time, distance and socio-economic status [6].

1.2 E-learning Development and Growth

Since 1980, the technology utilized to develop the learning methods. At the same time the dissemination of computers for personal use supported to conjunct the both to develop computer based learning. The higher learning institutions had efficiently changed their policies over the last decades, including wide range sharing, quality assurance and long life learning.

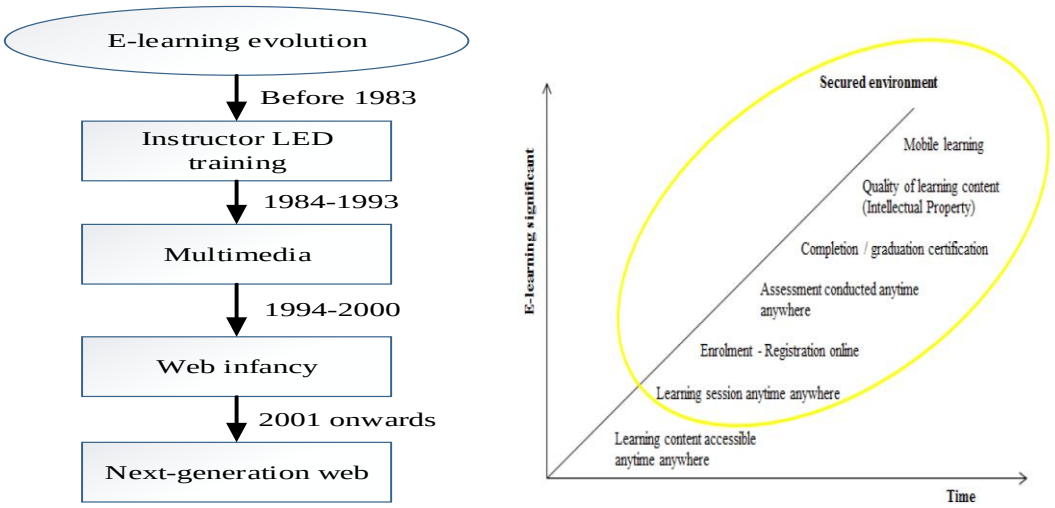

Fig 1.1 Growth of e-learning

Recently, the e-learning application is combined form of three approaches of technology utilization: asynchronously technology which is help or supplements face-to-face learning, utilizing synchronously and asynchronously technology as a help tool or supplement a face-to-face learning, the last one is asynchronously and synchronously technology utilized to distribute a learning course.

Fig. 1.1 illustrates the evolution of e-learning from the year 1983 to the present day. Before 1983 the instructor led training was the mostly used tool before the development of computer became available widely. In this type of education system, the students and instructor are seated in classrooms. Later the multimedia invented at the period of 1984 to 1993, windows 3.1 and CD-ROMs were the important advanced technologies at this period. The dynamic presentations and classroom discussions were the two sessions of learning method.

The electronic or technology developed and utilized in e-learning. There is a requirement to switch to the learning in assuring the e-learning success. The use of technology in learning includes distributed teaching, e-learning, distance teaching and online classes and also some general terms which are used as interchangeably. The related form of self-learning is distance education. In this type of learning, the education materials are given be accessed in online or can via mail. In some cases, the

meeting of instructor with learner also had few times per semester in distance education. But the online learning classes are popular recently [7]. Furthermore this development, a new trend in this technology started from the year 2008 arrived, where content include digital media like text, image, audio and video are through various communication standards supported by smart phones and different computing devices which can support mobile learning technology for distance learning [11].

1.2.1 Factors of e-learning development

The environmental nature including asynchronous or synchronous communication will produce a high-level interactive environment in e-learning that permits learners to share the information and describe how to recover the efficient information. In addition, the satisfaction of environment will improve learner's perceptions of technology that may advance their learning process participation. Furthermore, to share the knowledge and experience, the activities of learning in e-learning offer a suitable opportunity for instructors and learners. Basically, for developing an e-learning environment four major elements have been considered such as learning activities, learner's characteristics, environmental characteristics and environmental satisfaction.

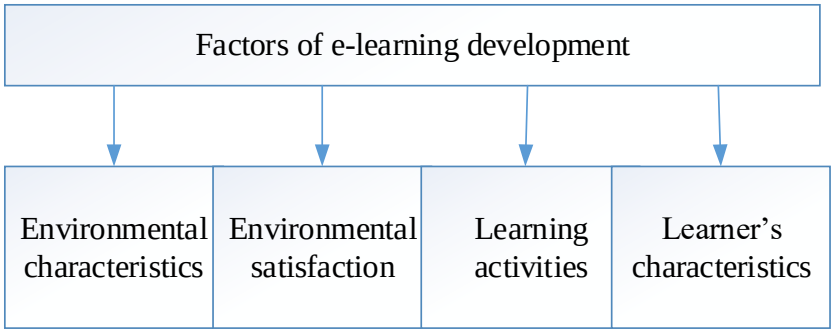

Fig 1.2 E-learning development factors

1.2.2 Considerations for developing efficient e-learning

Three considerations are suggested for designing efficient e-learning systems: learner's characteristics, instructional structure and interaction. The understanding of target

population is important for developing the e-learning. Initially, learner characteristics including self-efficacy, self-directed behaviour, and autonomy required to be recognized. The direction of multimedia activates learners to advance difficult cognitive skills including understanding main elements of conceptual difficulty, capability to utilize the required concepts for reasoning and inference. Based on the interaction, learners boost their communication with instructors and other learners, also they can increase their opportunities for constructing their own knowledge. Typically learning takes place within a social context and the process which can additionally develop the mutual construction of understanding.

Hence, three considerations in designing efficient e-learning environments with respect to the e-learning basic: learner's self-efficacy, multimedia formats and communication environments. For teaching and learning, understanding user's attitudes toward e-learning simplifies the formulation of suitable e-learning environments. Basically, using a single linear methodology the approaches of e-learning assessing cannot be established. Particularly, multidisciplinary approach is required to perform survey on individual attitudes towards e-learning. A useful diagnostic instrument can be designed based on the observations of learners which must incorporate the various attitudes of user perceptions

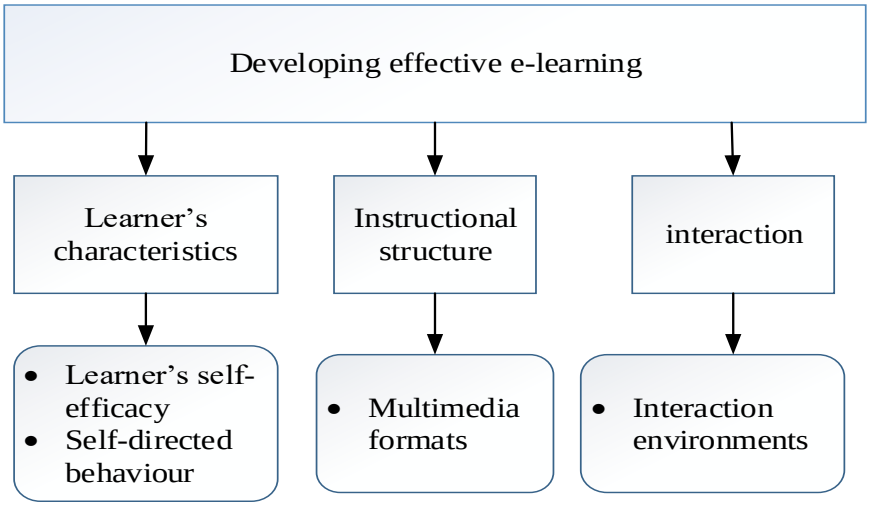

Fig 1.3 Developing effective e-learning

1.2.3 The Three-tier technology Use Model

The Three-tier technology use model (3-TUM) coordinates multidisciplinary aspects such as inspiration, social cognitive theory (SCT), theory of planned behaviour (TPB), and technology acceptance model (TAM). 3-TUM categorizes attitudes every individuals toward information technology as individual characteristics and system quality tier, the affective and cognitive tier and the behavioural intention tier.

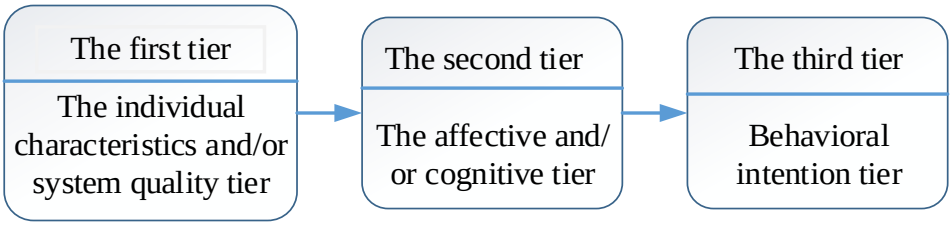

Fig 1.4 Three-tier technology use model

The individual characteristics and system quality tier is used to describe and to calculate, how individual affective and cognitive components will affect the individual system quality and properties. How affective and cognitive components influence individual behavioural desires examined by the affective and cognitive tier [8].

1.3 E-learning overview and structure

The e-learning is the methodology which is used to define the advanced internet and different web technologies, with the objective teaching and learning experience improvement. The e-learning systems are acting as enablers which strengthen their requirement correspond to educational techniques. The e-learning systems are invented instead of the traditional teaching and learning and also it enables various types of users to operate more services and materials. For defining instructional techniques, the e-learning success determinants' understanding it's important.

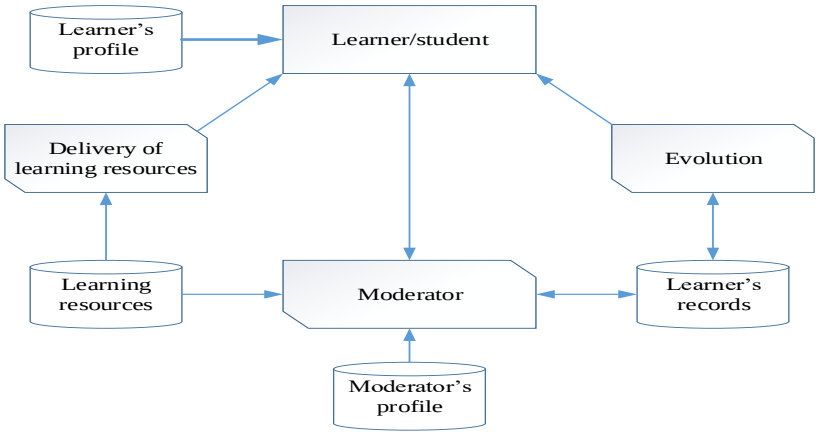

Fig 1.5 E-learning overview structure

The figure 1.5 describes the typical overview of e-learning system. The student accesses the resources and does the activities as instructed. The moderator plays as a teacher me calculate and plays as moderator of the learners. Finally, there is a calculation process. Fundamentally, this illumination consists of the processes and the databases.

The e-learning contains different classification of media which distribute text, images, audio, streaming video, and animation. It contains some digital implementations like satellite TV, video or audio tape, CD-ROM, and computer-based learning, as well as local extranet/intranet and web-based learning. Whether free-standing or based on either internet or the regional networks in networked learning, underlines several e-learning procedures in information and communication systems. The process of e-learning can happen in either classroom or out of the classroom. The learning can be either self-paced asynchronous learning or instructor-led synchronous learning. A flexible distance for learning is suitable for e-learning, it can also utilized in conjunction with face-to-face learning, in which case generally, the term blended learning is utilized.

1.4 E-learning system components

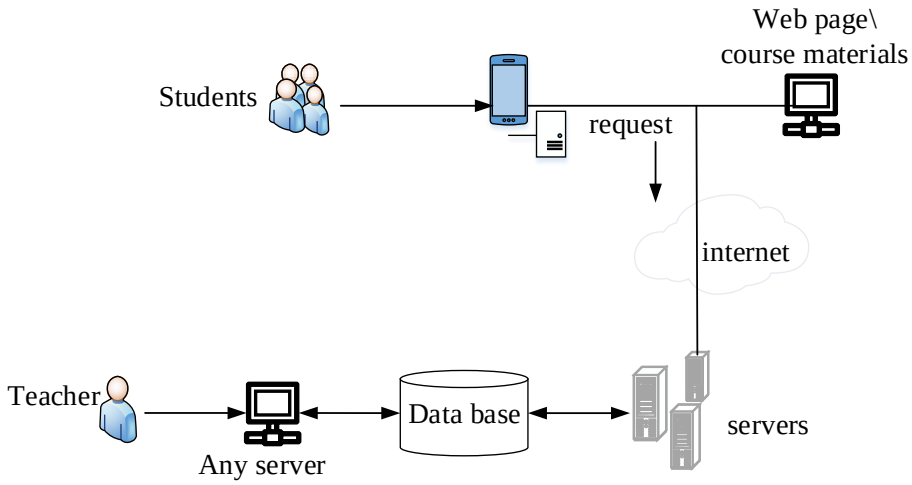

Fig 1.6 E-learning system components

The market of e-learning is still continuing to upgrade regardless of the claim that lots of e-learning actions have dropped short of expectations. During the fall 2007 term, the Sloan foundation reports indicate that minimum of 3.5 million learners were registered in one online course. This growth of e-learning is feeded by recent institutions entering into the online region, linked with a demand continuous students for online learning choices. The requirement of knowledge workers has also provided to increase the usage of e-learning: each employee needs to arm themselves with the knowledge to as a suit a degree as possible [9].

1.5 User expectations in e-learning

The relation of the students, education experience in instruction fields such as course structure, instructor and student's interaction, interaction with peer learners and individual learning capacity were analysed. The instructor supports in learning, it strongly correlated to course satisfactory and learning achievements. The teacher's counselling and providing are the most important for student's perception towards their construction of knowledge, acquisition of media competency and also the satisfaction towards the course.

Moreover, the learners highlight the teacher's knowledge in the application of e-learning courses. The e-learning instructors are facing extra tasks when compared to classroom discussions. So the instructors have to upgrade coherent and designed learning material which technically constructed. Then prepare the students to face self-tests or online practices and also collaboration online with peer students. In the case of blended learning, instructors need to combine face-to-face learning and online sessions, maximum as possible.

Table 1.1 Statistics of student's expectations

Contents
Importance of variables concerning course design -Fair and standardized construction and learning contents of the course -The platform utilization -Suitable cost-benefit-ratio of work and outputs of education Importance of variables regarding the discussion with the teacher -Quick feedback from the teacher -The instructor support for counselling and education -Chances to communicate personally with the teacher -Simple and rapid availability of the teacher -Knowledge of the teacher in the application of courses in e-learning Importance of variables regarding the interaction with peer learners -Simple and quick interchange of information and knowledge with peer learners -Different communication utilities for exchanging the information among peer learners -Support of combined learning and group work with other course students -Personal connect with every peer learners Importance of variables regarding individual learning processes

-Convenient of education concerned with place and time
-Convenient in choice and pace of education
-Chances for self-paced chapter practices and the implementation of one's knowledge
-Chances for controlling one's education outputs
-Helping inspiration

Importance of variables regarding learning achievements

-Gaining knowledge and techniques in the subject trouble
-Gaining skills on how to use the knowledge
-Gaining skills in communication and cooperation
-Gaining skills in self-regulated learning
-Gaining skills internet utilizing for scientific training

1.6 E-learning security risks and threats

In the era of e-learning development, highly concentrated on possible technicalities for providing and distributing e-learning content. But the security requirement in the environment of e-learning has often been ignored. Providing a secure end-to-end session between the learner and the institution's e-learning network were the main role of security, where security can be described by means of technical tools. This can be illuminated by the application of data integrity with the help of data encryption through virtual privacy for groups based on e-learning.

The problem of security within an e-learning environment is considered as another attention from the view of student. For providing student's interaction and collaboration, there must be need of concentration on constructing a sense of security. This encloses the requirement to give the privacy and trust for learners towards learning. The capability for a learner to manage a personal space is superior, particularly when the personal information is distributed which is necessary to preserve

privacy for students. The other one is trust which is an old age problem which is used to describe the trusted things. It means that the user feels secure on something trusted and is proved as trustworthy. In environment of online e-learning the trust is essential because, when an interaction in physical is rejected then the virtual interaction is the only option. When the mechanisms is placed to produce privacy and trust, then the students can feel more confident on collaborating and interacting with others through virtual communication.

The security were not implied in early development stages of the web-based application, later the security was added because of the real-life application scenarios such as e-learning and e-commerce. Many types of security issues in the web-based application are identical to ones of e-learning due to the correlation with e-learning. For the secure e-learning environment needs to ignore the four classes of threats: modification, interception, interruption, fabrication, and respectively. The recent researches on security e-learning mainly concentrated on three regions: policy, identity and intellectual property. The realization of threats or risks caused an amount of loss in the asset. Through the medium of vulnerability, the risks and threats are realized. The main threats and risks are given below,

- Confidentiality violation: a person who is unauthorized to have access, obtaining the access of the assets illegally from the E-Learning model.
- Integrity Violation: an unauthorized person accessing and tempering the asset utilized in e-learning model.
- Denial of Service: During the transaction among the users of E-Learning model, sometimes the legitimate access rights may be prevented due to disruption in the traffic.
- Illegitimate use: Exploitation of privileges by legitimate users.
- Malicious program: a sequence of codes for destructing the other programs.
- Repudiation: transaction of documents for participants may be denied by other persons.
- Masquerade: The truth hided by the way of behaving by the hackers.

- Traffic analysis: the information leakage due to abusing the communication channel.
- Brute-force attack: an experiment tries to uncover the right one, using all possible combinations.

These threats following risks may happen during the access of textual and non-textual messages among various participants in the e-learning model. The knowledge is situated on the internet and can be operated by every individual who learns from e-learning environment. Hence, security and payment problems become important and difficult and also these issues need to address properly.

In order to save the privacy perceptions of learners, an e-learning model must be highly secured against illegal temper or data abuse. It must have verification and authorization support for the users. Then, the electronic learning materials copyright and license agreements have to be considered to neglect the malicious use of data. For this purpose, encryption approaches and digital signature techniques have been utilized to save knowledge on the internet. Finally, when the electronic materials for learning are delivered to learners through the internet, the intellectual property rights of owners should be compensated. Hence, the system of e-learning has to support creation and management of the learner's account and gain online money transactions. And also it is necessary to construct and settle the internet accounting and billing services to satisfy all these requirements [11].

1.6.1 Author's risk

The authorization to access materials such as journals papers, books, etc. to a lot of friends, learners and connections for authors proved by advance technology implementation. The authors are the authority to develop and apply the contents. The students who already registered only can access the lecture notes, assignments, class test papers etc., In E-learning, the essential duty of authors is to ensure the safety of their own data in various aspects like unauthorized use, modification, reuse etc., Since these data can be destructed or modified by others called hackers via various attacks, it

is the duty of the authors to provide the assurance for the data accessed by the users are unmodified and also integrity of the text have to be checked. Apart from these security aspects, breakdown of some components like hard disk, network connections etc., must be considered which is being useful in risk analysis. Because failure in these hardware components cause loss of data integrity. Furthermore financial details play an important role generally useful in risk analysis.

1.6.2 Teacher's risk

The teachers have the duty to provide help in all possible directions for the students according to their academic matter. According to the need of the course, the teachers can follow their course material and presentations or buy from a third party. All e-learning risks can't be controlled to the practical system which is important to protect the whole methodology of examination, learning, grading and evaluation. There are different kinds of teacher, who are delivering the classes in different manner. But there are some common risks during the actions such as distributing lecture, sending assignments and notes, accepting and evaluating answer sheets, preparing and delivering mark sheets.

The interaction between the students and instructors are an important component of teaching any course. The online interactions have some advantages, since all the documents are saved on a server electronically. But interaction using digital storage system faces the risk of privacy of learners and teachers. In order to make the students to understand the concepts clearly, the students and teachers need maximum interaction. The robust security tool can only direct to this type of communication over a long period. Cheating in the examination is also considerable risk for the teachers. The teachers should be worried about the availability and non-repudiation of evaluations apart from cheating.

1.6.3 Manager's risk

The degree certificate is provided to a learner, after the course has done, by the concern board or Authority provides diploma in any E-Learning system. To set up and

run an e-learning institute, the board fixes some rules and regulations. If there is any doubt in those rules, it becomes difficult during the inspection time. The major risks in E-Learning are masquerading of persons to write the exam on behalf of the enrolled student and also supporting the students illegally during online examination. Additionally, some of the legal problems may be also a great risk for the managers like online testing, copyright, sharing the official files etc.,

The enrollment in a course is takes care by managers and the rejection of enrollment concerned when needed. Enrollment of any specific student in more than one course is possible and this will be risk for huge organization. In this case, a plan for backups and recovery process test has been required. Alternatively, during the time of need, it will be complex to make the data up to date. It is also risk for the management to deliver duties in sensitive problems such as managing password for each routers and servers, daily network traffic recordings, periodical power supply to the server and other network devices.

To control the authorization of access strategies for the learners is the duty of manager i.e., to allow the students and the other participants to read, write and execute, in order to run the system effectively. Otherwise maintaining the privacy is difficult. Therefore the manager essentially assigns a system of people and authorize the user of e-learning system by means of several operation such as index i.e., to permits deletion and creation of files, modification i.e., to grant deletion or addition of attributes in a relation, resources i.e., to permits creation of new tables, and Drop i.e., to permits deletion of relations. The e-learning materials bust be also protected from unauthorised processes like read, insert, update, and delete on components from databases of many kind.

1.6.4 Student's risk

The maximum numbers of users in the e-learning system are the students whose purpose is to acquire knowledge and to share their knowledge with others. The learners group can be categorized into various stages from down level, diploma, degree, post

graduate, up to doctoral level. But the user need to be more conscious about all materials obtained from institute, instructors or other learners. Otherwise the students will have to challenge the issues in examination if the attackers have modified the question papers or another main file. The learners must have the awareness against the malicious login information, because some intruders can prevent the authorized students using various attacks so that they cannot access the e-learning server for learning.

The students need to be trained to work individually without the support of teachers because teachers are not always available to support the students. The students required to have great writing and communication skill. The misunderstanding of what was meant is happen when the teacher and another student are not face-to-face communication. The student's feedback mechanism will always enrich the teacher; there is a difficulty from the learner's part to send the same feedback to the e-learning management. Finally, all users need to know about original e-learning websites, since the attackers can fix up the fake websites by phishing. Also the human eye cannot find the difference between original and attacker site easily. So the e-learners are incited to have some information to enter the original e-learning website, which must be highly confidential [12].

1.7 Need of e-learning security

The requirement of e-learning security at the time of including learning systems on the internet provides possibly many chances for efficiency improvement and reducing cost, also it provides possibly unlimited risks. The internet offers much higher access to data and more valuable data, not only to the legal users but also provides to the criminals, disgruntled employees, hackers, and corporate spies. The use of standard interfaces and protocols were increased and has offered important benefits for the user community; this also initiates access for an attacker.

The growth of using virtually standard databases, spread sheets and other generic software implementations and elements and of standard hardware processors combined

together with the continuing development and dissemination of hacking mechanism and methodologies creates the intruder's subsequent deeper intrusion into our information systems ever simpler. Moreover, it is difficult to find such attacks and tough to identify their resource and their safe location, hence such attacks are more tempting [13].

1.8 E-learning applications

The system framework has been utilized once by an organization, after the actual applications are identify that can be improved or acquired. The learning management system (LMS), on the basis of notation that content can be purchased while the organization has been relatively enamoured by LMSs. In this case of specific subject matters like IT training, the content needed for improving an organization and transfer of knowledge are unlikely to be identified from the third part vendor. These types of contents required to improve the particular needs or these needs to be modified for an organization to cater the language and cultural needs.

1.8.1 Learning design system (LDS)

To permit content developers to rapidly analyse and construct educationally sound learning programs is the main function of learning design system (LDS). LDS provides several choices of design methodology which will be incorporated to the project management capability. The learning objects contain this structure, later be utilized by content producers to improve educational resources. Some of the e-learning systems have not recognized because they do not have adequate instructional goals and objectives through which their complete potential can be described. These objectives can never describe beyond a wide statement of direction in the worst cases. The key idea to improve efficient material lies with combined form of clear learning goals and pedagogical design principles.

A project management tool is used in which the instructional design tasks are embedded where the tasks can be assigned and tracked. This technique is helpful for the developers to utilize and follow a suitable educational design procedure in making

education resources. Also this technique will permit content producers to plan and process the improvement of their own e-learning project when supplemented by templates.

1.8.2 Learning Content Management System (LCMS)

The key role of LCMS is to provide collaborative authoring environment in order to make and maintain learning content. A process of workflow that can enable coordination among the collaborative authoring within the LCMS. Matter Experts and Content producers use the LCMS to improve the content, while Media producers are used to add multimedia components and resources used for interaction. Lastly the writers use the LCMS to evaluation and support the submitted articles. Generally, the instructor would have processed such activities. To manage the process of content development the LCMS provides a structured framework where a group of people involved in the improvement process.

In distributed web applications, it is possible to create and assemble the content from any remote locations. The content developers can collaborate in a symmetric manner is provided based on revision tracking, task notification and check-in/check-out facilities. The main objective of LCMS is the production of educationally efficient learning content with in a fixed time and within the estimated budget. The gap between authoring tools and LMSs are connected by LCMS.

The LCMS offers the designer with the utilities and functionalities needed to create and efficiently address the learning content. The LCMS permits organizations to:

1. Capture the knowledge within their organization.
2. The knowledge structure concentrated in to direct learning programs.
3. Incorporate third party content.

Throughout the organization, the rapid updates, management, dissemination and employment of that knowledge are achieved.

1.8.3 Learning Support System (LSS)

The LSS is developed to support teaching and learning activities which is web based-environment. The LSS is a managing tool which helps to a group of learners from the perspective of an instructor. It is also a constructing to tool, which is useful to the instructor to construct learning materials. The syllabus consists of resources assembled from teaching articles produced in the LCMS and/or pages particularly produced in the LSS. Then the teacher can utilize the LSS to plan for distribution of the resources, whether through normal classes or electronically distribution techniques. Using industry standard protocols, the LSS shall offer following and employment of data to the LMS.

The LSS offers to the learners so that they can access the syllabus planned by the instructor from the perspective of a learner. The syllabus which relates the materials additionally contains a form of a study plan or a schedule of tuitions. The LSS should support the learners with suitable collaboration tools: synchronous messaging, threaded discussions, and shared whiteboards additionally to the learning materials. To assist the education procedure, the LSS assures a support to the user for class projects and assignments if necessary. A repository is provided for the learner to access materials which could be useful for students to share their knowledge. Also to distribute quizzes, tests, surveys, and other forms of assessment the LSS can also utilized.

The major features of LMS, has the tendency to support self-directed learning while most LMSs tend to offer these abilities. First and foremost thing of an e-learning methodology is to produce a social space, and manage it for, the teaching and learning requirements of the specific group of inhabiting that space. The requirements of the particular learners in the course, is a platform that can be simply changed to take into regard. E-learning can never apply that all students going to particularly evolving at their individual pace, fairly many organizations will resort to offer elearning help to general teacher led programs. Efficient e-learning programs will help to growth the

communication and collaboration between instructors and learners. The ideal class is arranged around the 50/50 rule showed from the studies by OTTER Group.

At least half of the learners in the virtual classroom spend the time to interact with other students to learn about them. Due to this student un-satisfaction occurs dramatically, when the social aspect of the classroom is missing [14].

1.9 Integrity verification

Integrity verification is a method ensures the integrity of media by detecting the tampering attempts towards the original content. Signal transformations affect the watermark which is designed with semi-fragile watermark. Tampering does destroy or alter the semi-fragile watermark, once the semi-fragile watermark is altered then it could not be used for authentication. The trust is important in the context of learning. Initially, the learners likely to believe that all sources of information they found and accept whatever they read as true, where integrity of content is being essential. Naturally the provider has the tendency to gain sufficient compensation for his investment, towards the construction and application of e-learning content in terms of time and money.

Therefore protection of e-learning content is being essential, in order to prohibit the unauthorized learners from accessing the e-learning system. A support of metadata is needed to increase data integrity, which is useful to document the information of learning pattern. Metadata helps the students to escape from false teaching materials.

The basic process in e-learning is teaching objects. The proposed approach can be described in two steps. In the first step, common parts of several lessons are generated to satisfy reusability and which can be used in all lessons, hence it makes one or more learning object. Then a watermark is embedded in all objects. Normally, information about application domain of learning object is included in the watermark. Then patterns are saved in a repository. A metadata is useful for describing any learning object. So in the second step, data of data doing watermarked in teaching patterns.

Let us consider a JPEG image as an object, which can be represented by its size and format known as metadata. This watermarked metadata can only decoded by the authorized learner in each objects. The authorized learner can also detect the changes made by the unauthorized users. Suppose that if JPEG image is converted to GIF image by unauthorised user, then the authorised user can detect these changes.

1.9.1 Integrity verification technique

The authenticity of data is described as integrity of data, which means whether the data has been tampered with. Recently two techniques used for digital media integrity authentication; cryptography authentication and digital watermarking technology.

In traditional cryptography authentication, all types of data are treated as binary bit stream. A message authentication code or digitally sign the hash value is produced with asymmetric encryption algorithm. This message entailed at the end of the real message and sent them out together. Authentication failure could be directed by any change of bit in the received message. The level of modification and location cannot be determined by this cryptography authentication method but it can provide high level of security.

For authenticate the integrity of the content, the digital watermarking technology is embedded to determine the author and buyer of the content. This type of watermarking is required for limited applications, which need data integrity. The digital watermarks can be split into two; robust digital watermark and fragile digital watermark. A robust digital watermark is useful for copyright protection and fragile digital watermark has been utilized for limiting the tampering event depending upon usage. The robust digital watermark is complex, it is difficult to remove, highly resistive to original data process or malicious tampering. There will be some dissimilarity between extracted fragile digital watermark and the real watermark.

This property is useful in identifying about tampering in order to recognize whether the data is authentic. Therefore a fragile digital watermark is also used to authenticate the

digital media and provide integrity protection of the same. Fragile digital watermark has two advantages; the first one is, it never requires any additional authentication information to be attached at the end of the original data media. The second is, the watermark can be shared in all parts of the digital media which achieve the security. Moreover, a fragile watermark can easily locate modification and point out the extent of modification. It can also recover the original data from the modified data. [15].

1.9.2 Watermarking

Watermarking is a data-hiding technique where a message which has confidential information is hidden into audio clip, an image, video sequence or any other work of media. Recently digital watermarking is one of the emerging research area, which mainly focus on copyright protection, used in wide range of applications. Based on the application needs, different categories of watermarking schemes have been designed. Applications like ownership identification and copyright protection generally needs a robust watermark, which can withstand attacks such as common image processing operations. Another category of watermarking namely, fragile or semi-fragile watermarks which are mainly suitable for content authentication and integrity attestation. These watermarks can be fragile to attacks, where changes in an image and also changes of localization are possible due to attacks.

Watermarking is categorized in to three domains: integrity, copyright watermarking, and annotation watermarking. Copyright is useful to provide security for owners, which is applied on copyrighted resources, to know the copes are authentic or tampered and evaluate copyrighted multimedia data. These functionalities can be achieved by analysing the spread spectrum of the data over networks and servers. The goal of Integrity watermarking is detect any changes in the content, through embedding some integrity information to the media so that the digital content is being protected from illegal activities. Annotation watermarking embeds some supplementary information to the media directly, so that it cannot be separated accidentally from the media. It is also known as caption watermarking. With the aid of

Sajjadi et al. [16], this thesis aims to design an integrity watermarking to provide authentication for learning contents on an e-learning system.

1.10 Integrity Attacks

An active attempt to customize destruction of knowledge in the e-learning without accurate permission is known as integrity attacks. The modifications including changing, creating, and removing the data of data and data are called integrity attacks. The granted users able to reach sources with integrity attack but they may find something irrelevant to what they expect.

1.10.1 Malicious Code Attacks

The unintentional causes may destruct or change the information by integrity attacks. However the malicious integrity attacks are becoming more general and harmful. The malicious code gets in different kinds such as virus Trojan horse, worm, etc. The administrators and end -users of the e-learning system have to secure and periodically check their systems to ensure that they are malware-free.

Message Injection Attacks: An attacker may freely inject the malicious messages to the e-learning system, via authentication attack, which may be seen as legal system traffic by the users at the other end.

Traffic Modification Attacks: The attacker may delete or reconstruct particular bits in the packet data and forward it as if the data never changed. This attack does not require knowledge of key data or understanding about the e-learning data itself.

Traffic Deletion Attacks: This is similar to the previous attack, which may simply delete the data on the communication channels. This also requires no specific understanding of the e-learning data itself.

Traffic Mirror-Rerouting Attacks: This attacker may reroute the traffic to unauthorized users. Hence the traffic may be destined to the authorized users and also unauthorized users found by the attackers. The attacker can distribute the mirror of the traffic to anywhere, hence it is difficult to detect this attack.

Traffic Miss delivery-Rerouting Attacks: Some or all end-users lost some messages are not received the messages if the attacker reroutes traffic to unauthorized receivers without governing. This attack is more recognizable than traffic mirror re-routing attacks. Since the end-users and messaging protocols can identify lost images.

Forgery (counterfeit) Attacks: This type of attack, make wrong representation of data that has come from another address. By interrupting a communication session an attacker can hijack a session and maintaining it in the name of the victimized end-user.

Stack Overflow Attacks: In order to increase the space allocated and spill over into adjacent data or code areas, the attacker intentionally provides a huge amount of input data (for example, 3000 characters in a limited range), either corrupting other values or inserting new commands to be executed [17].

1.11 Advantages and disadvantages in e-learning

1.11.1 Advantages

E-learning researchers made a number of techniques and identified several advantages in contrast with traditional classroom learning. Some of them are listed below.

Time and location flexibility: By the just in time and on the-job learning cancels the barriers of distance and time in e-learning, and has ability to attain a worldwide audience including disabled, part-time and non-traditional people.

Time and cost saving: Nearly 40% of money shall be eaten up by travel cost with traditional learning. E-learning significantly reduces the travel expense and time of travel, since e-learners never need to travel to a particular learning location. In general, the online training institutes are expecting and average of save cost around 40% to 60% and save time to 50%, compared to traditional education.

Self-paced and just-for-me learning: Due the activities of learner-centric structure, e-Learning is said to be self-directed and self-paced. Selection of teaching activity is based on the learner's background, interest and career at the moment that must be the highly suited to them. It does not depend on the information from passive receptor.

Hence this type of learning is highly effective and active participation of learners is increasing day by day.

Collaborative learning environment: E-Learning forms an online collaborative learning community, which connects the experts and learners who are physically separated. An e-Learning system motivates the students, to query any type of questions and elicit their individual opinions without inhibition. They can also distribute their own ideas with others in the learning system which is easy through online forums. These facilities are not possible with traditional classroom learning.

Better access to the instructors: Learners can obtain any kind of online guidance and support from instructors. More opportunities for communication can be perceived by the learners which is not possible with conventional classroom learning.

Unlimited use of learning materials: Learners in e-learning system are provided with information and knowledge 24 hours per day and these are kept always with high-quality and maintained safe. Therefore people are allowed to review the current or past information/knowledge stored in repositories at any time. Also the user can access and retrieve the learning materials again and again.

1.11.2 Disadvantages

Presently several e-learning systems are available. Each of which having their own demerits. Based on that, greater efficiency and higher societal potential of e-learning are hindered. These issues occur in present e-learning systems are given below.

Text-based learning materials: Some e-learning systems have only text-based content i.e., the screen contains large volumes of text. This will never engage the students effectively, during online learning.

Not enough content for good understanding: E-learners must be provided with adequate materials for proper understanding of the subject matter while many of the e-learning systems fails to do so. For example, some of the e-learning system uses the power point slides regarding subject lectures. This is not sufficient for the learners,

since the power point slides prepared by authors may not be understood by students. In this case students need some more material including synchronized instructional videos, additional lecture notes etc., to compete with traditional educational systems.

Less interaction and user flexibility: The involvement of online students in learning is a major issue, due to the separation of instructors and students by time and location. Whereas traditional classrooms provides face-to-face interaction with instructors and classmates. Recently, some of the e-learning systems are not support good interaction with the students. Hence these systems are less flexible due to the control over learning content and need more process in order to achieve individual requirements.

Unstructured and isolated multimedia content: The multimedia instructional materials can be easily uploaded through the web without additional processing in most current multimedia based e-learning. They are generally proposed in static and unstructured manner which do not have strong association among related content in different media [18].

1.12 Problem statement

The digital multimedia contents are slowly changing their traditional media counterparts in today's environment. This modification brings out easy to modify, edit and abuse the digital formats. Through internet and other communication technologies wide range of digital contents including e-mails, e-library, e-learning, e-commerce, e-books, news and Short Messaging Service (SMS) are used in live and business applications. The evolution of these digital text contents can tamper by malicious attackers.

In digital era the media contents are facing various problems including content authentication, tampering attacks, integrity attacks and copyright protection. The researcher's already concentrated on the content authentication and tampering detection for image, video and audio. But many studies ignored the text based content authentication and tampering detection. Hence the content authentication, tampering detection and integrity verification must get attention. Several techniques are proposed

to provide copyright protection, authentication and tampering detection for the digital text documents. Cryptography, steganography and digital watermarking are the techniques which utilized to avoid text document issues. Among these techniques the digital watermarking technique is suitable for the above mentioned issues in text documents.

In comparison with other watermarking techniques the text watermarking is difficult because the text can easier to delete, copy, modify and tamper. The zero watermarking is new technique which proposed as practical technique for authentication of text. In this technique the original content is not modified rather it uses the content of text to produce original watermark information. This study presents new approaches on the basis of content authentication and tampering detection of text documents. These techniques create the watermark data and embed it within text document logically without modifying the contents of the original text document.

The study develops and proposes a framework of Ngram order of Markov model and HSW based zero watermarking technique. This technique describes the major changes among the Ngram orders and their accuracy robustness, capacity and complexity. This study also discuss about the effects of the proposed technique on the text documents, volume and types of tampering attacks.

1.13 Work background

Although I started my work using text watermarking for Arabic linguistic before changed to English because the Arabic source in internet is limit scope and most of elearning books and journals in English language, where 1.60 million English papers published worldwide each year and 2.0 million English books published worldwide each year, and I have faced the data set problem where Arabic script has many features like points above or under its letters, kashida (extension character), and existing of diacritics. Any characters can take different shapes and non-vowel Unicode is different. And there are different Shapes of some Arabic letters where different write for same character in Initial عـ Medial ـعـ and Final ع and Arabic alphabets contains

Pointed ش and un-pointed Arabic letters like س, Arabic has 15 pointed characters from 28.

Where the alterations in contents like position of diacritics and/or characters are not acceptable. Moreover the approach in text- watermarking has not taken much concern as it is obvious from the overview on techniques of zero-watermarking.

1.14 Objectives of the thesis

This thesis mainly focused on the integrity verification and tampering detection of text documents based on watermarking techniques. The different techniques in watermarking are analysed in the presence of different types of attacks and threats. The objectives of the thesis are mentioned as below

1. To study e-learning systems and their applications and advantages in e-learning with the aim to analyse the main challenges and threats occur in e-learning.
2. To investigate and evaluate the different watermarking techniques such as zero watermarking, text watermarking, structure based and text content based watermarking techniques with the purpose to understand their features and prepare the ground for comparison.
3. To propose a novel tampering detection method for text document watermark protection.
4. To implement the proposed method in simulation environment to analyse the detection accuracy based on different attacks like insertion, deletion and reordering attacks.
5. To perform comparative analysis with the existing techniques to analyse the advantages and disadvantages of the proposed scheme.

1.15 Motivation of the thesis

The text watermarking method is important approach to avoid the issues and attacks in text documents. Several attacks are raises in text documents such as illegal copying, modifying and redistributions, the watermarking approaches will prevent form these

attacks. Several kinds of watermarking techniques are improved in the recent researches. However, the existing watermarking methods are not completely provides good performance in tampering detection and integrity verification for text documents. To resolve this issues advance techniques are required. The tampering detection and integrity verification performances of text document is required to be improved from the existing works based on the watermarking methods which is applicable to text documents. This thesis is formulated to prevent the text documents from various attacks like insertion, deletion and reordering attacks in text documents.

1.16 Contributions and Novelty

- NLP have been combined with ZDW in authentic and novelty approaches to improve robustness and security troubles for English text watermarking.
- Techniques of text analysis are deployed for watermark generation and extractions for providing solutions that are safe and complex.
- Avoid attempts tempering from all types of attack
- Watermarking is blended in a unique way which offers a robust text watermarking solution against attacks of various length and characteristics.
- Zero watermarking techniques adopt a novel approach for protection the sensitive content to embed a watermark logically where the content is not modified.
- This study can be used not only for e-learning systems but also it can be used for another e-services as e-library, e-government, e-banking, e-mails, and e-commerce.

1.17 Organization of the thesis

This thesis concentrated on the watermarking techniques to improve the integrity verification and tampering detection of text documents. The organization of the thesis describes as follows.

In chapter 1, introduction to e-learning, e-learning systems, barriers and advantages of e-learning are provided. E-learning security threats and issues also provided in this chapter.

In chapter 2, literature survey for the e-learning systems with different watermarking approaches for integrity verification is provided.

In chapter 3, text watermarking and integrity verification is mainly concentrated. The biometric cryptosystem with error correcting output code matrix is deals in this chapter.

In chapter 4, combined Markov model and zero watermarking technique for integrity verification of PDF English text documents is described and this method compared with existing methods.

In chapter 5, the proposed HSW based zero watermarking technique is presented and compared with existing method.

In chapter 6, the general summary and major conclusions of all chapters are provided with conclusion and future scope.

CHAPTER 2
LITERATURE SURVEY

CHAPTER 2

LITERATURE SURVEY

2.1 E-learning system

In e-learning systems, the learning styles which referred to student's preferred ways to learning were played important role. The system provided valuable advice and instructions to students and teachers with the knowledge information of various styles. To overcome the disadvantages of the traditional detection method, e-learning system permits electronically and statistical algorithms gave the opportunity. These types appealing reasons had directed to developing many studies focused the integration of learning styles and adaptive learning system. Huong May Truong et al. [19] focused on this reasons and studied 51 cases which carefully investigated into various sections of the integration process. It captured a different aspect from learning styles theories selection in e-learning environment, online learning styles predictors, dynamic learning models categorization to enormous learning models implementations.

The result of this work provided visions into various improvements, accomplishments and opened issues in the area. The article also provided conclusions, discussion and plans for future researches on the basis of these findings.

In information technologies, the e-learning emerged increasingly more essential for academia and teaching company and had become one of the most important advances and applications. Said S. Al-Gahtani et al. [20] adopted a quantitative method looking for an affective description of the decision personality's behaviour toward the assimilation and acceptance of e-learning in the settings of academic environment. This work conducted survey, had 286 participants to collect the research data. The framework of this study were on the basis of third form of the Technology Acceptance Model and the data were investigated by using structural equation modelling for describe the reasons that influences the learner's motive to e-learning utilize. The result of this work showed the predicting factors of e-learning technology acceptance, at the same time

some related post-implementation interventions also examined and expected to contribute to the acceptance and assimilation of e-learning systems. Moreover this work indicated and provided a direction to understand the success factors.

In education environment the use of e-learning technology is really identified as essential topic. The existing studies were implemented to know how efficient e-learning systems and useful. Nouzha Harrati et al. [21] conducted an experimental based research to investigate how lectures communicated with an e-learning environment on the basis of predetermined task form defining interactions in low-level. From the institution materials from the Computer Science and engineering departments, the client-side log data is gathered. Later, data were analysed to infer the usability degree from the estimated usage metrics together. The simulation results were explained that the system usability scale score were not enough judges to evaluate the correct satisfaction and acceptance level of instructors for utilizing the e-learning models. The presented methodology in this work had form towards evaluation of usability to acceptance development and experience of user for learners and educational staff.

The research community trusted e-learning ecosystem were the next generation of e-learning. However, the existing e-learning ecosystems lagged improve of basic facilities, which can automatically allot the needed computation and storage resources as services. Therefore, Bo Dong et al. [22] introduced a new methodology that cloud computing into an e-learning ecosystem. This work presented an e-learning ecosystem based on cloud computing structure. In e-learning ecosystem the cloud system structure and related tools allowed for the efficient materials use, equilibrium, sustainability and stability of an e-learning ecosystem.

In education process the e-learning systems were the enablers, which establishment their value as part of the learning technique. The factors of e-learning achievement were understood for teaching techniques. Various researchers had studied e-learning implementation and adoption and also several studies had addressed e-learning success

from various angles. But none of the research had verified whether student's cultural characteristics including individualism versus collectivism, played a crucial role in the success of perceived e-learning. Manuela Aparicio et al. [23] proposed a success model of e-learning systems such as cultural design, individualism/ collectivism.

This work also provided a clear knowledge of the effect of student's artistic nature, for individualism/collectivism, based on the understanding output of e-learning models used. This work reported experimental case improved through an automatic survey delivered to higher education learners belonged with several teaching levels and institutions. For obtain the results, this study implied quantitative methods. The result of this work were demonstrated the crucial role of individualism/collectivism on organizational and individual effects. Based on the collective culture perceive more organizational and individual effects than the individualistic culture students influenced. The result of this work showed that for the learners with a powerful individualistic culture, satisfaction played a central role in the way they evaluated the individual effects, and individual effects on organizational effects. This work discussed implied in real time applications.

Many studies showed that the e-learning can afford feasible and capable education because of its reliable and ubiquitous characteristics. The traditional learning methods limited by space and time limitations. The e-learning system were eliminated these problems by support a pervasive teaching environment. Beulah Christudas et al. [24] had proposed a developed method for customising education article for specific students from huge amount of database in e-learning model. This study adopted the compatible genetic algorithm (CGA) which developed the proficiency of the genetic algorithms by produce compatibility in the education patterns.

The presented work in this paper improved the quality of individual teaching procedure in an adaptive e-learning model and provided maximum suitable content for specific students. This work depicted the outputs of customizing the teaching procedure by the help of tuned compatibility level of the teaching objects with respect

to the education model. The forcing compatibility into the search space helped to reduce the search space and it filled the search space with suitable chromosomes. The experiment outcomes of this presented work provided development in marks of the students and also in their satisfaction levels. The proposed algorithm in this work were matched with the normal algorithms and showed better performance in fitness values and running time, number of executing generations. The overall outcomes of this work determined that customization of content transfer on the basis of behavioural traits of learners directed to great learning.

In skill training and knowledge acquisition, the web-based e-learning system (WELS) had emerged. Generally, many pre and post-adoption jobs relevant to estimation passed from the view of technology. Initially, the users had recognized commonly as presence a group key of stakeholders in effecting the implementation of information technology, their outlooks for this model were crucial. Daniel Y Shee et al. [25] had produced different standard technique from the viewpoint of student satisfaction to help those assessment based actions done in pre and post-adoption stages of the web based e-learning models phase.

The proposed methodology adopted based on the study of many standards decision making and with the perception of user satisfaction. Additionally, based on the proposed methodology, this work experimentally analysed students' opinions of the related concentrate of decision standards. The analysis done in this work handed a study of learners and obtained information were analysed based on the analytic hierarchy process. Finally, this work recognized that students considered the student interface as being the very significant measurement of decision standards.

2.2 E-learning system challenges (barriers)

In Yemen, the higher education eLearning system faced several challenges as the Ministry of Higher Education doesn't recognize and attest certificates from this type of learning, although there is an e-library development project that is supported from the World Bank.

Moreover, Ministry of Human Resource does not neither consider elearning certificates for any promotion for the government employees nor accept these types of certificate for employment process. Furthermore, the internet facilities at Yemen are too poor.

In Taiwan's higher education faced several challenges in the Ministry of Education loosed its control over degree-awarding programs. The research on stakeholder perceptions towards the essential e-learning problems became critical at this condition for the strategy producers to create feasible investment decisions toward e-learning programs. Even though, a methodology of instructor insight with equivalent techniques in e-learning had considered, the learner outlook lost in the instructor-learner dynamics. Yu-Hui Tao et al. [26] had developed a learner methodology on the basis of typology and questionnaire items utilized in existing studies about teachers, to understand teacher and student perceptions on organizational responsibilities towards the problems of e-learning. Two completely distinct groups of students: sceptics and the optimists were recognized in the cluster investigation. Based on the four higher-level problems have designed from a aspect study of 30 variables. Comparison results of teacher methodology and learner typology were concise along with techniques in the interaction and applications.

The implementation of electronic learning systems trained the users in advance technologies, services and products with the rapid change in all type of environments. For technology implementation and management, the huge investment in e-learning had produced user recognition as a critical matter. Chorng-Shyong Ong et al. [27] had proposed an advance design, seeming reliability, to validate the applicability of the technology acceptance model (TAM) in clarifying engineers' decisions to agree e-learning, and addressed a pragmatic technology administration matter. Even though, the user acceptance obtained clearly expanded in prior search, efforts were needed and validated former outcomes, mainly in various technologies, user populations and organizational contexts. From six international companies 140 engineers were taken as

sample and examined. The results of those sample results greatly supported the extended TAM in predicting engineers' intention to utilize e-learning.

In e-learning the success had more important because an unsuccessful effort to apply e-learning clearly reflected in terms of the return of investment. E-learning required careful consideration of the underlying pedagogy or how learning takes place online, which the most important prerequisites. But these methods were neglected in most of the works done in implementation of e-learning. Thavamalar Govindasamy [28] had focused to identify the pedagogical principles underlying the teaching and learning activities that constitute efficient e-Learning. In this work effective learning management system were presented. Many studies had conducted to failures from traditional and online learning courses.

Many works does not provided complete and reliable determination of dropout from academic courses. But this work had proposed a clear and exact definition of dropout from academic courses in the context of e-learning courses. In addition, those data were saved in literature that attended learners of e-learning dropouts at considerably most rates than their counterparts in tradition lectures. This work investigated to main designs: academic locus of control and students satisfaction with e-learning. The experimental outcome showed that the key indicator in learners' decision to dropouts from e-learning lectures were the learners' satisfaction with e-learning. In addition the results of this work showed that the academic unit of manage appeared to have no effect on learners' decision to remove from e-learning classes.

E-learning had utilized as an efficient solution by most organizations provided learning on demand opportunities to individual employees to decrease cost and time. The research guided to evaluate the achievement and efficiency of e-learning models in administrative environment had low because of the information systems success models attention among the researchers. Whether the basic data models success forms were continued to investigate e-learning systems success had scarcely resolved. Yi-Shun Wang et al. [29] had designed a multi-dimensional model for assessing e-

learning system success (ELSS), from the attitude of learner. The steps utilized in conceptualizing an ELSS construct, producing things, gathering data and calculating a multiple item scale for measuring ELSS were explained. Based on the analysed data from a sample of 206 respondents, this work had presented indication of the reliability, scales factor structure, convergent validity, content validity, criterion-related validity, and discriminant validity. The experimentally calculated instruments were utilized by researchers to improve and assess the theories e-learning models.

The problem of e-learning continuous utilization had main task in training and academia at the time of the growth of adopting the e-learning. Nuan Luo et al. [30] focused to get how interactions encouragement learners' intellect of community and continuous use objective of the e-learning environment. In this work 643 students were participated totally. The results of this work indicated that the interaction between student-instructor and interaction between student-student were particularly strengthened learners' sense of association and impact, and promoting their interests in the e-learning platform. Additionally, the relationship between interactions among members and their sense of community were moderated by student-content interaction. The results suggested that managers were required to provide relevant mechanisms to appreciate the interactions takes place in e-learning platform for the managing learners' continuous utilization, for online learning institutes.

For self-learning, with the diffusion of easy-to-use web 2.0 tools including podcasts, blogs and wikis, e-learning had become a widely used tool. The individuals utilized these mechanisms, in the plan that improves their training. Soheila Mohammad yari et al. [31] had proposed that an individual's level of digital literacy disturbed the performance through its effect in the performance and effort expectations. Using e-learning and their performance were continued to explain the influence of digital literacy on the intention of individuals. Later, the concept of digital literacy integrated with the unified theory of acceptance and also it tested with proposed model using survey data. The results of this work indicated that electronic learning on operators' performance and cared out opportunities, effectiveness anticipations on continuance

target, and continuance objective on efficiency, these relationships were significant. These findings in this work suggested that individual elearning assists the utilization of e-learning, and it intentioned when studied the effect on rendering.

2.3 Integrity verification and watermarking

The improvements in communication and information technologies were provided new ways to access, share, manipulate and replicate the medical images. Gouenou Coatrieux et al. [32] had presented a medical image integrity verification system to find and exact local malevolent image alterations. Based on the non significant region watermarking with signatures, the proposed integrity analysis process extracted from different pixel blocks of interest and which compared with the recalculated ones at the verification stage. A set of three signatures were produced in this work. The initial two sets were devoted to detection and modification location were cryptographic hashes and checksums, during that time the final set issued from the image moment theory. In this work, initially showed that how geometric moments were utilized to approximate any local modification by its nearest generalized 2D Gaussian. Later, the ratios between real and recalculated geometric moments can be utilized as image features in a classifier based technique in order to describe the nature of global image processing were demonstrated. The experimental results of this work considered both local and global modifications in Magnetic resonance imaging (MRI) had retina images illustrated the overall achievements of this technique.

Dual watermarking combines the robust as well as fragile watermarks into the same cover image. With this scheme, both copyright protection and integrity verification can be achieved. Most of the other work state of art approached either lacked the feature of tamper detection and real content recovery or provided an approximation based on the adoption of coarser block level approach. Priyanka Singh et al. [33] had proposed self-recoverable dual Watermarking technique, which integrated the functions of tamper detection, copyright protection and recovery into one scheme. The proposed methods were independent of the order by which the embedding of robust and fragile

watermarks, as these are embedded in various locations of the cover image. This technique performed both tamper detection and recovery at the pixel level. This proposed method obtained recovery information for every 2×2 image block in simple eight bits additionally encoded through recording table to only four bits.

The decrease in recovery bits allowed effective embedding of copyright information which analysed against comprehensive set of attacks. This method identified to be robustness against errors such as noises, motion blur, filtering, histogram equalization, rotation, jpeg compression, etc. The evaluation of the extracted copyright information done by the help of various objective error matrices, depending on the correlation between pixels, their locations and values separately. The proposed method of this work tolerated tampering ratios up to 50 percent through the visual quality of the recovered image deteriorates with the increment of tampering ratio. On the basis of normalized cross correlation comparative results, the probability of fault acceptance, the probability of the wrong elimination and peak signal to noise ratio metrics validated the efficacy of the proposed scheme of this work over other existing state of art methods.

The improvement in web technology and availability of various resources such as images, e-journals, e-books, videos and another Medias increased the requirement of copyright protection of all electronic information. Due to the increased data in internet were led to risk in several methods such as digital data can be copied and republished by any other name. Based on the development in computer technology the digital data were easily modified and manipulated. For multimedia contents over the internet, digital watermarking affords validation, authentication and copyright protection. In addition to audio, images and video clips, the text forms were the most widely utilized communication. Therefore, the text must be protected.

The watermarking technique had developed and protected from illegal copying, imitation, and prevents copyright violations. Leena Goyal et al. [34] had proposed an algorithm that assured the confidentiality and integrity of the file. In this proposed

technique the watermark were produced on the basis of contents of the document and embeds it without made any altering in the contents of the file and also encrypted the scrip to afford confidentiality. The watermark easily extracted and verified for tampering to authenticate and prove the integrity of the document. The experimental results of the proposed work showed best performance compared to other works.

Protection of copyright and authenticating digital contents were the main problems in internet. The technique named a digital watermarking provided complete authentication and copyright protection. In internet, large amount of video, image, audio and text contents were handled in internet and needed complete protection. Zunera Jalil et al. [35] had done by combined image plus text watermark a novel text watermarking algorithm. Based on this technique the text document protected completely. The techniques watermarking the text were developed to protect the text contents from threats such as copying, redistribution and copyright violations. The proposed watermark technique in this work logically embedded in the text and then extracted later and proved ownership.

2.4 Zero watermarking

The technique which named zero watermarking helped to avoid the threats like copyright protection. Yaxun Zhou et al. [36] had proposed a zero watermarking based on combined form of discrete wavelet transform (DWT) and singular value decomposition (SVD). The proposed scheme in this work constructed a watermarking from the image properties which extracted from the original image by the DWT and SVD the watermark embedded, instead of changing the original image information. Initially, based on the DWT the original image decomposed into the appropriate levels, and the gained approximation images were divided into non-overlapping blocks. Later, the SVD applied to the each block to get the singular values. At last, the embedding of zero watermarking realized by the help of XOR operation between the initial singular value of the each block and the pixel value of the certain binary character watermarking basically. The proposed zero watermarking pattern in this work, ensured

the watermarked image quality without any distortion and resisted the common image processing attacks including noise addition, filtering, compression and geometric cutting by good robustness.

The internet and other communication technologies were globally developed. Due to that reproduce, disclose and distribute digital contents were very simple. Additionally the profits of information interchange obtained. The digital community challenged through authentication, copyright protection and forgery threats. Zunera Jalil et al. [37] had proposed a zero watermarking algorithm on the basis of occurrence frequency of non-vowel ASCII words and letters for copyright protection of clear content files. The occurrences of non-vowel ASCII letters in content partitions were incorporated by embedding algorithm and formed a key based on the watermark. The water mark extracted from the noisy text to identify original copyright author by extraction algorithm. The results of this work proved the efficiency of the algorithm adopted on text encountering dispersed deletion and insertion attacks happened at random.

In the community of digital world, the text watermarking technology for copyright protection had a major interest and many research works had played effective roles in digital text protection. But, the works had low practical value because of the pertinence to literary form. Meng Yingjie et al. [38] had designed a zero watermarking scheme for copyright protection based on the keyword set, the core verb set and the proportional feature of adjectives as the watermark information elements based on the characteristics of prose. This work also established the construction and the detection process model, designed algorithms of the main modules. The experiment results of the proposed scheme in this work showed great robustness and false rate positive rate.

To solve the contradiction between the watermark embedding capacity of and invisibility, Gui Feng et al. [39] had proposed an improved zero watermarking algorithm. Many works based on zero watermarking methods were utilized and without the need to see the embedding size. The procedure of the proposed zero watermarking algorithm: initially, the text carrier translated into the respected binary

image, later the watermark constructed by using the DCT combined with the logistic mapping, finally the similarity criteria were utilized for the watermark detection. The performance of the adopted algorithm in this work showed the robustness, invisibility and the capacity of watermark, after several attacking observations.

In modern world the authentication of medical images for telemedicine implementations were the important problems. The medical field challenged of accurate consistency and verification of medical images occurred as major matter. Anushikha Singh et al. [40] had proposed the zero watermarking method. The proposed method solved the problem of medical image protection for several applications. This proposed method in this work provided protection for medical images without tampering the medical image. Domestic properties in the Singular value decomposition (SVD) domain were utilized and digital binary code (master share) generated for each fundus image. The master share dramatically joined with encrypted patient ID resulting into a secret share. The patient ID exactly recovered by the authorized person only on access of the produced secret share. The proposed watermark method in this work analysed on the publically available DRIVE dataset of fundus image and results accomplished were encouraged in the way of medical image identification and verification. The work also utilized in telemedicine implementations where perfect and loss-less identification required for medical images.

The technique of digital watermarking helps to protect copyright and ownership of digital information. In traditional digital image watermarking methods, the texture of original image gets distorted more or less. [41] This article proposed, a copyright protection scheme based on zero watermarking is introduced. Instead of embedding the watermark in to the host image, it is encrypted with the host image. Two approaches have implemented which are using Discrete Wavelet Transformation and Singular Value Decomposition to extract robust features of host image. In approach number one, the host image is divided in to overlapped sub images, each with size of 8×8 pixels. Then Discrete Wavelet Transform is used followed by Singular Value Decomposition to decompose further, each block up to level-one.

Whereas in the approach no two, initially the image is first subjected to DWT, then approximated and later it is divided into overlapping sub-images of size 4 × 4. The further steps are similar to the previous approach. The secret image with the host image is encrypted by generating two shares: the master share and the ownership share. The master share is generated by differential classification of features extracted. The ownership share is generated with the help of master share and secret image. An individual share cannot give any clue of the secret image, when these two shares are operated together, they can reveal the encrypted secret image. The performance of the proposed approach towards resisting image processing and geometrical attacks was explained via simulation.

The explosive growth in network and multimedia technology, had possibilities of the digital media re-transmitted or duplicated throw net. This introduced a big challenge to protect multimedia copyrights. The technique called digital watermarking implemented to rectify this challenge. Hung-Hsu Tsai et al. [42] had proposed a zero watermark technique with geometrical invariants by the help of support vector machine (SVM) classifier against geometrical attacks for image authentication. The studied method also called as SVM-based zero-watermark. The proposed scheme made no change to original images during the embedding, the owner signature of the images achieved high transparency. The RST invariants of images discovered based on the discrete Fourier transform (DFT) with log-polar mapping (LPM).

The secret key generated by the proposed method in this work for a host image through performing logical operation exclusive disjunction, an exclusive or (XOR) operation, on the original watermark and a set of the features of the RST invariants of the host image. Subsequently, a trained SVM (TSVM) is regarded as a mapping so it memorized the relationships between the set of features of RST invariants and the secret key. Initially, the TSVM obtained the estimated secret key by fed with the set of characteristics of RST invariants of the watermarked image. After, the SZW method extracted the estimated watermark by performing the XOR operation on the set of characteristics of RST invariants and the estimated secret key. In this work, the particle

swarm optimization (PSO) algorithm also adopted for the purpose of search for a set of almost best parameters of the SVM. Where experimental outcomes of this work showed better performance than the other works.

2.5 Text content based watermarking

The digital contents were generally comprised of text, video and audio. Copyright protection and authentication of digital images, audio and video were proposed by existing studies but authentication and copyright protection of clear content had discounted. Maximum digital contents including e-books, news, articles, chats and SMS were in the formula of original content. In digital contents, the Copyright protection and authentication were the important problem in digital epoch with efficient communication mediums including internet. Zunera Jalil et al. [43] proposed a novel zero-watermarking algorithm for authentication of plain text.

For information exchange, plain text widely utilized medium utilized over the internet and the verification of authenticity of information were important. For plain text watermarking and authentication, had limited techniques only available. This work presented a zero water marking algorithm which generated a watermark based on the text contents. Later, this watermark extracted by utilizing extraction algorithm and proved the authenticity of the text file. The experimental outcomes showed the performance of the proposed algorithm against tampering attacks. Also watermark accuracy and distortion rate of 10 various text samples have been verified with varying length and attacks.

There was a limitation on adequate techniques for tamper detection and content authentication of plain text based on digital watermarking concept. To achieve content authentication and tamper detection for English text documents Fadl M Ba-Alwi et al. [44] had proposed an intelligent text zero watermarking approach on the basis of probabilistic patterns. The letter based on Markov model of order THREE, abbreviated as LNMZW3 designed for analysing the text and utilized the interrelationship between contents of given text to produce the watermark in proposed approach of this work.

After, the watermark extracted by extraction and detection algorithms, the status of text document was identified that whether it is authentic or tampered. For implementation of the proposed approach in this work utilized the programming language, PHP with Net Beans IDE 7.0. Moreover, this proposed approach detected the tampering attacks irregularly even the volume of the tampering were low, medium or high.

Huge text data including articles, letters and documents were shared in web due to the quick improvement and broad application of internet which directed to the copyright protection challenge. For digital copyright protection, the digital watermark utilized as the important technology. But many works presented based on the text watermarking and only few works focused the zero-watermark. Yingjie Meng et al. [45] focused for promote the Chinese digital scripts copyright protection technique, this work proposed a Chinese text zero-watermark scenario on the basis of sentence's entropy. This opinion computed the entropies of sentences based on word frequency and created essential selection on the basis of entropy. Later, the watermark designed with the order of important sentences. Moreover, the proposed scenario were validated by simulation and proved its effectiveness, robustness and the capability to prevent the attack.

The copyright protection and authentication are very important for digital contents, since the internet has been growing day by day in terms of speed, storage, number of user etc., The existing researches on watermarking schemes for text content highly focused on the copyright protection but had low consideration on content authentication, tamper detection and integrity verification. Fahd N Al-Wesabi et al. [46] presented a novel English text zero watermarking technique on the basis of probabilistic patterns. The core algorithm of the watermark generation was presented by the Hidden Markov model along with watermark detection.

The proposed Hidden Markov model was designed for text analysis and watermark detection on the basis of correlation between contents of host text document. A two

dimensional matrix was used for this purpose which contains probabilistic weights of states and transitions. The watermark extracted by extraction and detection algorithm. Based on that the status of the text content including authentic, tampered, integrity, etc were identified. Moreover, the success and feasibility of the presented approach in this work were proved with algorithm experiments and relative with other approaches under irregular deletion and insertion attacks. The experimental results of this work showed that the propose approach obtained more security and better robustness.

In digital watermarking of content media, the contradiction between the invisibility and robustness were considered to be more important, for short redundancy on the text carrier, which complex to use to the needs of watermarking algorithm. Hence the research community searched for new techniques. Gui Feng et al [47] proposed an advanced zero watermarking algorithm which solved the contradiction between the watermark embedding invisibility and capacity. The existing works considered so many watermarking techniques and without requirement to embedding capacity and robustness consideration which best than the traditional watermarking algorithm. This study zero watermarking had some steps: at first the text carrier were translated into the responding binary image, later combine DCT with the logistic mapping utilized and constructed zero-watermark, the correlation characteristics were utilized for the watermark detection. The performance of the algorithm showed the watermark capacity, robustness and invisibility of the algorithm after many attacked experiments.

The purpose of digital watermarking is to provide copyright protection and authentication for digital contents in web. The internet contains the plaintext in the form of image, audio and video. Also the content of research papers, newspapers, messages, e-books, and articles were the some examples of plain texts. Hence, the plain texts required full protection. Through the digital watermarking, the authentication and copyright protection for digital contents over the Internet can be obtained. The textual contents were the main components of the internet. Therefore, the protection for plain text documents required huge attention. K.U Jaseena et al. [48] proposed an invisible watermarking algorithm where combined form of text and image

watermark for text document protection was utilized. The design was used non-vowel ASCII characters for watermarking. Here the watermarking key was generated end embedded in to the text logically. After the watermark extracted and proved the identity. The simulation experiments of this work showed that the proposed method were efficient in order to text files integrity.

The benefits of quick and easy digital information exchange over the internet had attracted issues and attacks in the form of digital attack which affect the information integrity, protection and authentication. Due to the influence of textual-data sent online, these issues were more outstanding for text-content. Omar Tayan et al. [49] proposed two robust zero-watermarking approaches: method A and B were capable to detect any content modifications. These methods avoided any modifications/embeddings on the real text to be published. This work addressed mechanisms for assuring intact integrity and authentication of samples of the sensitive textual content published over internet through the help of zero watermarking.

The presented approaches provided suitable indication of the relative sensitivities of every approach to third party alterations at the time of key-extraction phase. Proposed article included a compared analysis of the proposed approaches against already existing appropriate state of the art strategies on the basis of two cost functions were particularly useable to the objective of this work domain with suitable outputs. For authentication purposes, the derived systems were capable of achieving a pivotal requirement by assured textual context could be traced back to its real distributer. In the case of third party modifications the document tampering had detected.

In text document, the content authentication had main concern in digital environment. Xiton Qi et al. [50] had proposed Chinese contents authentication by zero-watermark algorithm. Initially, the frequencies of various part of speech (POS) tags were gained through natural language processing technology. Later, utilized to compute the suppose value and entropy, had text features. By one dimensional forward cloud model generator based on the expect value and entropy, watermark were generated. In

the trusted third party named certificate authority (CA), which watermark registered and stored. In this work the comparison between the disputed content watermark and its watermark registered in CA were calculated if authentication required. The experimental results showed that the presented algorithm were robust against content preserving attacks while sensitive to malicious tampering.

The watermarking methodology achieved the copyright protection of the multimedia contents. While multimedia represented various media including text, audio, video, image, and graphic objects, and they declared very various features in hiding information included, several watermarking algorithms were suitable to the multimedia contents and they developed. Based on the patterning the inter-word spaces, the text documents were watermarked. Young-Won Kim et al. [51] had produced text watermarking algorithm in this work. The proposed algorithm exploited the novel concepts of word categorization and inter-word space statistics. The words were categorized based on some properties. Various adjacent words were grouped into a segment, based on the help of word class information the segments were categorized. In each segment classes, the same amount of information inserted. Some statistics of inter word spaces of the segments belonging to the same classes were modified, based on that the information were encoded.

In internet, the texts were the important medium. The text watermarking techniques had improved for text protect from unauthorised copying, redistribution and to prevent copyright violations. For multimedia contents in web, the digital watermarking provided authentication and copyright protection. K. U Jaseena et al. [52] had proposed a new text watermarking technique which utilized the combined text and image watermark and encryption to the protection of content. The watermark logically embedded in the text and encrypted. After the encryption, the text decrypted and the authenticity proved by the watermark extraction.

To protect the copyright from the copying and tampering, the text watermarking provided feasible functions. Ping Zhu et al. [53] had studied a text zero watermarking

algorithms on the basis of the relation between the Chinese characters and phonetic alphabets. The proposed algorithm in this work extracted the features of the plain text by valuating based on the custom of Chinese phonetic scripts corresponding to the predefined interval threshold. After the chaotic transformation, the text features embedded with watermarking information in the Chinese typescript. The results of this work showed that the possibility of tampering was limited up to 0.1%, with the proposed watermarking algorithm and hence proved that this algorithm was strong in robustness and resistance to aggressive behaviour of attackers.

CHAPTER 3
TEXT WATERMARKING AND INTEGRITY VERIFICATION

CHAPTER 3

TEXT WATERMARKING AND INTEGRITY VERIFICATION

3.1 Introduction

Recently, the information in digital world has been attaining high growth. The computers and networks handles and stores high amount of data compared to hard copy formats. The electronically stored text documents sometimes goes under the attacks such as unauthorized copying, tampering, redistribution of the copyrighted contents, and illegal authentication. There are different keys like as authenticity, integrity, copyright protection, and confidentiality are utilized to avoid these threats. By the help of watermarking methods all of these threats can be solved. Lots of watermarking methods were improved and utilized in this field. Some of the schemes concentrate on the updating of graphical characteristics of litters, words, or lines etc. and another scheme concentrate on syntactic or semantic based techniques.

The unauthorised reordering and malicious attacks were increased rapidly based on the simple and online accessible of several software and techniques. Therefore, strategy protecting the multimedia content including cryptography, digital signatures, steganography, watermarking are developed and serve as present research areas [54]. The information stored or sent, should show up accurately on received at the other end of communication network, for the integrity. Finally, for non-repudiation, the sender should not be able to reject after sending a message at the time of having done so. An advanced technology named digital watermarking, must essential in place to secure the integrity or digital information. The watermarking technique inserting the information into a cover messages and after extracted for several functions such as identification, authentication and verification [55].

3.2 Integrity verification process

The integrity verification is intended for checking the integrity of an object. During the communication the content may go through low levels of compression. Therefore the system must be robust against low levels of compression. Any type of modifications made on the image had to be measured as a malicious attack, and has to be identified at the receiver as a tempered image, since there is no modification in the image is possible [56].

3.3 Classification of watermarking techniques

Recently, many advanced watermarking technologies developed for copyright protection, content authentication and integrity verification.

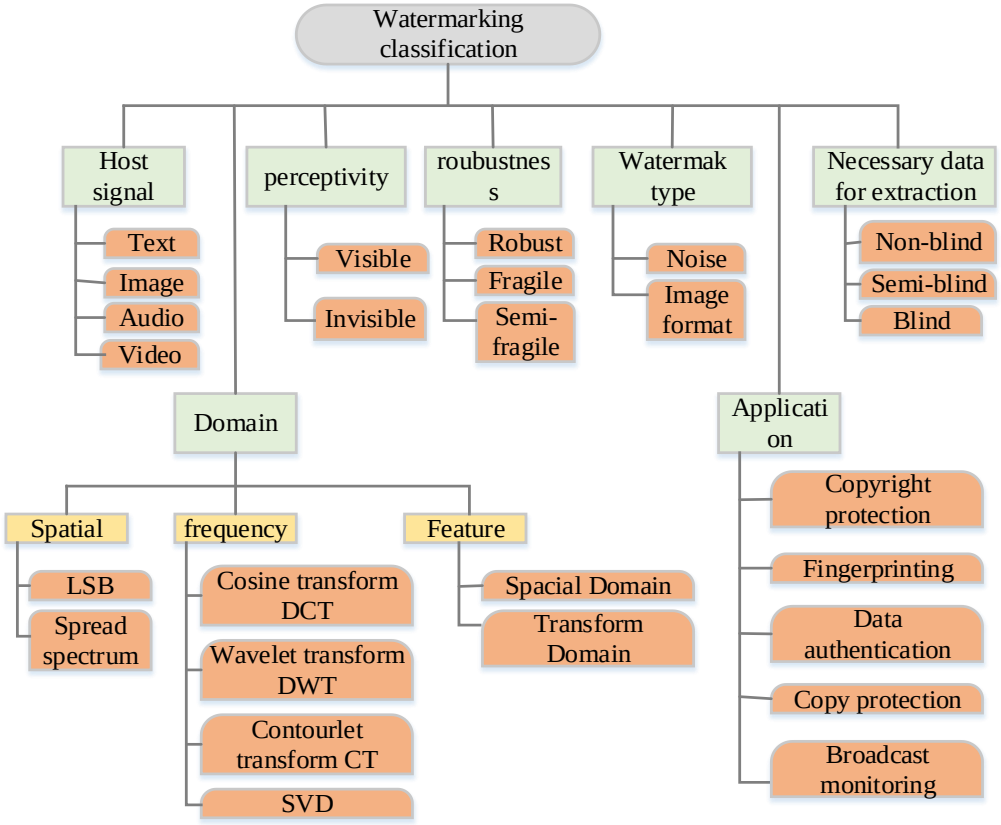

Fig 3.1 Classifications of watermarking

Fig 3.1 describes the classifications in watermarking. The watermarking technique categorized based on the domain, application, host signal, perceptivity, robustness, type of watermark and necessary data for extraction respectively. The host signal categorized as text, image, audio and video based watermarking. Visible and invisible watermarking techniques are perceptivity type. Robust, fragile and semi fragile watermarking techniques are utilized for robustness. Blind, semi-blind and non-blind are utilized for necessary data for extraction. Spatial based, frequency and feature based techniques are categorized under domain.

3.3.1 Text watermarking method process

The text watermarking method adds watermark to the PDF, DOC and other text file to secure the modifications done to the text. The text watermark applied in the font shape and the space between character and line spaces.

3.3.2 Audio watermarking

Many watermarking schemes are concentrates on image and video watermarking. Only few techniques reported in audio watermarking. The process of embedding a watermark signal into audio signal is called audio watermarking. The sensitivity of Human Auditory System (HAS) makes the audio watermarking as difficult process. Further, Audio watermarking techniques can be categorized based on the domain where the watermarking takes place.

Time Domain Audio Watermarking: watermark is directly embedded into audio signal in the time domain watermarking techniques. In this process, no domain transform is needed. Before embedding operation the watermark signal is shaped to assure its inaudibility. Time domain watermarking systems utilize various techniques to increase the robustness of the watermark. Initially the original audio signal is modulated and then filtered by low pass filter. Filtering reduces the distortion caused by embedding the watermark. In another way, the actual audio signal is subdivided into segments, then each segment is watermarked individually by embedding the same

watermark. HAS masking effects are used to shape the watermark signal. Shaping process is done in frequency domain while the shaped watermark is embedded into audio signal in time domain.

Frequency Domain Audio Watermarking: To embed an inaudible watermark signal in digital audio, the audio watermarking technique will be performed in frequency domain, and then HAS is used for masking of audio characteristics. To embed the watermark into perceptually important components, the transformation of audio signal from time domain to frequency domain activates the watermarking system. Initially the input signal is converted to frequency domain where the watermark is embedded. To gain the watermarked signal, an inverse frequency transform is applied to the frequency domain. There are several methods to embed the watermark in to frequency domain components. The techniques are mainly depends on the concepts of spread spectrum communication. A narrow band signal transmitted over the high bandwidth signal which creates them undetectable as the energy of the signal is overlapped. In this way the watermark increases over multiple frequency bins so that the energy in any one bin is very low and surely undetectable.

Wavelet Domain Audio Watermarking: Based on the use of wavelet transform, the signal decomposed into two parts, low frequencies and high frequencies. The low frequencies part is decomposed again into two parts as low and high frequencies. In this process, the number of decompositions is generally determined by application and length of original signal. The gained data from the total decompositions are called the DWT coefficients. Moreover, the original signal can be redesigned from these coefficients. This redesigning is called the inverse DWT. In wavelet domain, a method of audio signal watermarking uses patchwork algorithm. A binary watermark is embedded one bit in one data block in this method. For the purpose of robustness, watermark bits are locally repeated. At the point, where the watermark is going to be embedded, a number of bits are added in front of watermarks bits, called as synchronization bits.

3.3.3 Image watermarking

In spatial domain or in terms of frequencies in transform domain, images can be represented as pixels. We use reversible transforms like Discrete Cosine Transform (DCT), Discrete Wavelet Transform (DWT), or Discrete Fourier Transform (DFT) to transfer an image to its frequency representation. By modifying either the pixel values or the transform domain coefficients, the watermarks can be embedded within images.

DCT Domain Watermarking: The high frequency components are watermarked in frequency domain. Initially the image is divided into square blocks of size 8×8 for DCT computation in this method. A mid-frequency coefficients pair is selected for modifications from 12 predetermined pairs. Various methods are utilized for these mid-frequency coefficients. A new method modifies DCT coefficients satisfying a block site selection constraint. Certain blocks are chosen on the basis of Gaussian network classifier decision, after dividing the image into blocks of size 8×8. Later, the middle range frequency DCT coefficients are customized, utilizing either a linear DCT constraint or a circular DCT detection region. Most of the frequency-domain algorithms make use of the spread spectrum communication technique.

DWT Domain Watermarking: recently, the transform has been globally studied in signal processing in general and image compression in particular. The wavelet based watermarking schemes outperforms DCT based approaches in some implementations. The image decomposed into three spatial directions by the wavelet transform. Therefore, wavelets reflect the anisotropic properties of HVS more accurately. In the lowest bands (LL), the magnitude of DWT coefficients is larger at each level of decomposition and is smaller for other bands (HH, LH and HL). The detection of watermark at lower resolution is computationally effective because few frequency bands involved at every successive resolution level. Watermarking in the DWT domain consists of encoding and decoding parts. Initially, an image decomposed into wavelet frequency domain to get decomposed image in the encoding part. Image permuted watermark we add to gained image decomposition. The watermark

permutation is reversible for exact watermark extraction and it is the key. The two dimensional (2D) inverse DWT (IDWT) considered in decoding part.

DFT Domain Watermarking: By researches the DFT domain has been explored because it provides robustness against geometric attacks such as rotation, scaling, cropping, translation etc. there two type of DFT based watermark embedding techniques: in is directly embedded technique and another one is template based embedding. By modifying the phase information within the DFT, the watermark is embedded in the direct embedding. To estimate the transformation factor, a template is a structure which is embedded in the DFT domain. Once an image is transformed to a template, which is utilized at the detector to extract the embedded spread spectrum watermark [57].

3.3.4 Video watermarking

Robust invisible video watermarking techniques varies in terms of their capacity, the domain in which the watermark is embedded or detected and the real-time performance. The degree to which all three aspects are incorporated and their resistance to significant types of attacks are also being essential to assure robustness. Based on the domain in which the watermark is embedded, the methods can be divided into three major groups. According to the dimensionality of the transform, the transform domain techniques can be further subdivided. These classifications are helpful since the constituents of each leaf group share similar nature. Most of the presented video watermarking schemes are on the basis of techniques of the image watermarking and applied to raw video or the compressed video. As some problems in video watermarking is not present in image watermarking, like as video object and redundancy of the high amount of video data. To develop various schemes, researchers have make use of those characteristics [58].

3.3.5 Perceptivity type watermark

Visible watermark: These types of watermarks are accurately seen by the viewer and also recognize the logo or owner. The original signal is changed in visible watermarking technique, computationally these are low complex. The watermarked image cannot with stand the signal processing attacks, since the watermark can be cropped from the watermark image by possible attacks. The suitable option is spreading the watermark throughout the image, which protects the image from illegal authentications, but the quality of the image is degraded.

Invisible watermark: These types of watermarks cannot be seen by the viewer or owner. When compared to the original signal the gained output signal does not get much change. The watermarked signal is usually equal to the original signal. The imposter cannot crop the watermark as invisible watermarking as the watermark is invisible. Invisible watermarking is more robust against several signal processing attacks matched to visible watermarking. Also it does not change the image quality much, hence it is suitable for all the applications [59].

3.3.6 Robust type watermark

Watermark robustness determines the ability of the watermark to be hidden which ensures the legitimate regular usage or image-processing manipulation, including intentional or unintentional attacks.

The intentional attacks targets to destroy the watermark, while unintentional attacks do not explicitly intend to modify it.

Some of the common modification such as unauthorized removal or alteration of the embedded watermark and unauthorized embedding of any other information are referred as intentional modifications. There are some unintentional modifications due to image-processing operations including filtering, scaling and compression. Usually robust watermarks are useful for copyright protection to determine rightful ownership.

Semi-fragile watermarks are aimed to detecting any illegal modification, In other words, semi-fragile watermarking techniques can identification common image-processing and content preserving noise like lossy compression, from unauthorised content tampered.

The purpose of fragile watermarks is to detect the unauthorized modification. Fragile watermarking techniques ensures complete integrity verification. The fragile watermark can be changed or damaged by a slight modification of the watermarked image [60].

3.3.7 Application based watermarking types

Copyright protection: For copyright protection the visible watermarking is utilized, which is the most essential digital watermarking implementation. The embedded watermark cannot be avoided without data distortion because of the requirement of high level robustness in copyright protection. The visible watermarking is the most important digital watermarking implementation which is utilized for copyright protection. With a minimum data distortion, the embedded watermark can be removed because the copyright protection needs huge level of robustness. Later, this watermark is extracted to prove the true ownership, when someone claims the ownership of the data. Digital watermarking can be utilized to securing the redistribution of copyrighted material over the un-trusted network like the internet. Moreover, at the time of sharing the data over the internet, a secured way is required for distribution. These watermarking techniques have several applications.

Finger Printing: the finger print is similar to serial number to any product. Every distributed multimedia copy is embedded with a distinct watermark. The concept of the fingerprinting is utilized to find the exact owner of digital content. Each and every consumer of digital content has its own identity as fingerprint.

Integrity protection: the invisible watermark is like a proof of ownership. The main aim of this application is to discover the modification data. To analyse the authenticity

of received data, the data watermark is embedded in host data. In this case fragile digital watermarking algorithm is needed. For detecting the tampered regions and estimating by how much and how the data has changed, fragile watermark technique is utilized.

Broadcast Monitoring: broadcast monitoring is needed to secure the valuable TV products like news contents, from illegal transmission. This technique performs cross-verification, to find whether the broadcasted content that was supposed to be broadcasted has actually been broadcasted or not. The digital watermarking technique can also be utilized for broadcast monitoring. It supports an important application; commercial advertisement broadcasting, where the advertising entity can able to monitor whether their adds was really broadcasted at the right time and for right duration [61].

3.4 Stages in watermarking

3.4.1 Watermark embedding process

In this process in grouping stage watermark image and text file will be given as input and base on this input image to text converter will give the output as a watermark alphabet (WA). That watermark alphabet, preposition and group size will going to generate MOFL list and base on this watermark key generated. As shown in figure below.

The algorithm which embeds the watermark in the text is called embedding algorithm. The inputs for this algorithm are the text document and combined image and text watermark. A pre-processing on image and the text is required to convert the watermarks, pure alphabetical in nature. Fig 3.2 describes the Embedding process in text document.

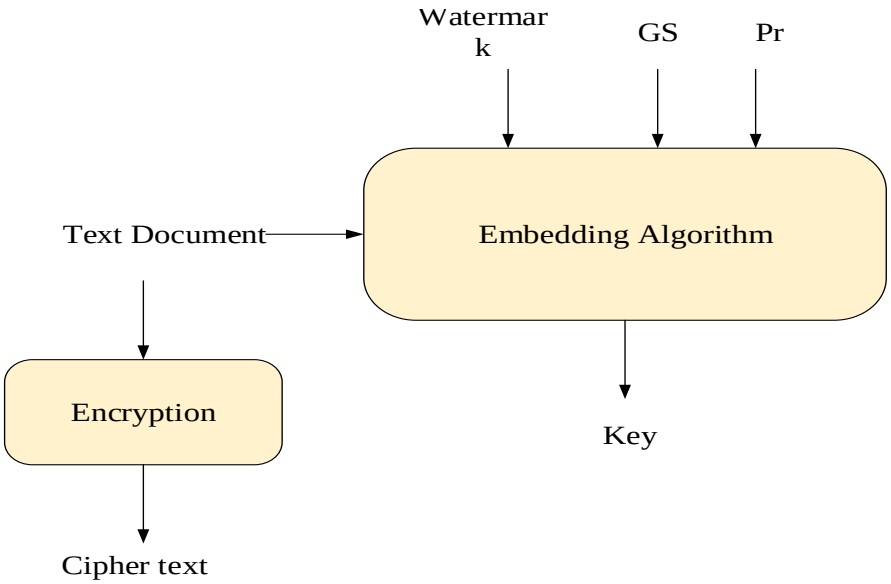

Fig 3.2 Embedding process

1. An image and text watermark and text document are combined and taken as input.
2. Separate text and image watermarks form the combined watermark.
3. Pre-processing the text watermark in order to make the watermark pure alphabetical, which removes special characters, white spaces, digits etc.
4. Pre-process the image to convert the image into gray scale and scaling to standard size (100 x 100 pixels).
5. Convert image to plain text by normalization process.
6. The two textual watermarks (watermarks obtained after text pre-processing and image pre-processing) and partial key containing a partition size (Pr) and group size (GS) is given as input to the embedding algorithm.
7. Generate the watermark key using the inherent properties of text by embedding algorithm.
8. Encrypt the text document using RSA encryption algorithm to increase security of text which is presented below.

3.4.2 Watermark extraction process

In this extraction process watermark key and text file will be given as an input base on this double letter processing will be there and MOFL list is generated than watermark generator going to generate watermark alphabet (WA). Than text to image converter will output watermark image. Figure 3.3 describes the extraction process of the watermark in a text document.

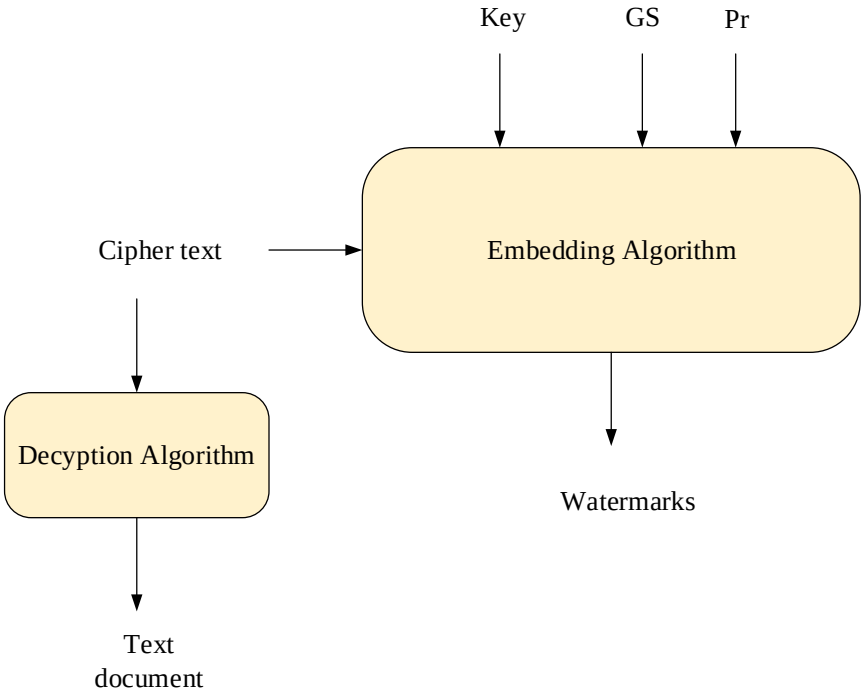

Fig 3.3 Extraction process

The algorithm which extracts the watermark is called extracting algorithm. It extracts the watermark (image and text) from the watermarked text using the key and watermarked text as inputs.

1. The inputs are watermark key and encrypted watermarked text.
2. Decrypt the text to recover the watermarked text.
3. Partition the text using Pr from watermark key.
4. Combine the partitions based on GS to form text groups as in the case of embedding algorithm.

5. Identify the occurrence of double letters in each group and second maximum occurring letter (2MOL) in each group.
6. Generate watermark from the text using extraction algorithm, using the contents of watermark key [62].

3.4.3 Watermark Generation process

Let the image in spatial domain be indicated by $f(x,y)$ with $0 \le x \le N, 0 \le y \le M$ where image size is $N \times M$. When the image is transformed into frequency domain using two-dimensional Discrete Fourier Transform (DFT), let it be denoted by $F(u,v)$ and the watermark which is utilized in the DFT domain expressed as $W(u,v)$ where $0 \le u \le N, 0 \le v \le M$.

Choose the watermark Logo such that multiple watermarks can be embedded into an image.

$$size|W(u,v)| << size|F(u,v)| \tag{3.1}$$

The watermark logo is an original valued Fourier domain matrix and the matrix consists of very high amount values for the Logo and zeros for the background:

$$W(u,v) = \begin{cases} 0; & (u,v) \in \{background\} \\ M; & (u,v) \in \{"Logo"\} \end{cases} \tag{3.2}$$

It is necessary that the chosen Logo is symmetric in both x and y directions and multiple Logos are same to one another. This constraint minimizes the distortion due to the conversion from spatial domain to FFT to spatial domain and does not degrade the detection.

3.4.4 Embedding algorithm

Let the watermarked image and a segment of the original image in the DFT domain, be indicated as $\hat{F}(u,v)$ and $\hat{F}_i(u,v)$, respectively. This scheme needs that each segment of image $F_i(u,v)$, to be of the same size as the watermark $W(u,v)$. Then the watermarked image in DFT domain is defined as

$$\hat{F}(u,v) = \sum_{i=1}^{P} F_i(u,v) + W(u,v) \tag{3.3}$$

Where, image is partitioned into P segments.

3.4.5 Detection of watermark

An algorithm designed to extract the watermark from attacked signal called detection (often called extraction). An algorithm can easily detect the watermark from the watermarked content, which is being unaltered during transmission and the watermark is still present. The extraction algorithm should be capable to perfectly produce the watermark in the robust watermarking applications, even if the modifications were strong. If any modification made to the signal, then extraction algorithm should fail in fragile watermarking.

This process is somewhat vary from the other watermark detection methods since, it merely tries to reveal whether an image is watermarked or not. The major objective is not to extract a watermark explicitly through this new scheme is capable of retrieving one without the possession of the original image. The suspected watermarked image in the DFT domain is segmented into P blocks which do not overlap and are superimposed on one another avoiding the segments which carry the strong features of the original image. Usually, the strong features of an image in the DFT domain are contained in four comers of the image unless the transform is shifted to the centred. The watermark size or the image partition size to be chosen such that excluding four blocks from the four Comers would eliminate the strong contribution from the image

in order for smooth detection of the watermark. Therefore, the above process can be expressed as,

$$\hat{W}(u,v) = \sum_{\substack{i=1 \\ i \neq q}}^{P} real\left(\hat{F}_i(u,v)\right) \tag{3.4}$$

Where, $\hat{W}(u,v)$ is the recovered suspected watermark whose dimensions comply with the expression and q signifies the blocks, which contains the strong characteristics of the original image, The recovered DFT domain matrix $\hat{W}(u,v)$ may create a watermark when plotted if the image is watermarked and a background noise if not. This result is to be expected as the watermark embedded was a real matrix in the DFT domain and the less contribution from the real image adds up to background noise in the process.

3.4.6 Detection algorithm

Once the suspected image is confirmed to contain the watermark, it can be completely recovered from the image by using the original image. The process can be defined as

$$W_c(u,v) = \hat{F}(u,v) - F(u,v) \tag{3.5}$$

Here, the watermark collection $W_c(u,v)$ has the same dimensions as of $\hat{F}(u,v)$ or $F(u,v)$ which contains multiple identical watermarks [63].

3.4.7 Possible attacks

Generally an embedding algorithm accepts the host and the data, and produces a watermarked signal. Then watermarked signal is either transmitted or stored. This digital content may be modified by a person, ie., the content is said to be attacked. Watermark attacks are intended to detect the specially crafted piece of data present in the watermarked content, without knowledge of the encryption key. To develop better watermarking techniques, a special attention has to be paid to these kinds of attacks.

Watermark attacks are categorized into four main groups:

Simple attacks: These attacks can modify the whole frame hence damages the watermark. It does not put any effort to identify and isolate the watermark. Examples include frequency based compression, cropping and correction and addition of noise.

Detection-disabling attacks: These attacks attempt to make detection of the watermark impossible by breaking correlation. Geometric distortion like zooming, shift in spatial or (in case of video) temporal direction, cropping or pixel permutation, rotation, removal or insertion are used for these attacks.

Ambiguity attacks: These attacks produce fake watermarked data and embeds several additional watermarks, so that the detector is not able to determine which of the watermark is being the first and authoritative watermark. This will discredit the authority of the watermark.

Removal attacks: The removal attacks exactly estimate the watermark, separate it out and discard only the watermark. Examples are collusion attack, denoising or exploiting conceptual cryptographic weakness of the watermark scheme (e.g. knowledge of positions of single watermark elements) [64].

3.5 Overview of Text watermarking technique

Any books, article, newspaper, documents, and website are created by plain text. Internet medium also constituting plain text, which can exist in all the components of websites, e-books, e-mails, and SMS. Plain text is exposed to many intentional and unintentional attacks. Therefore, it is a vital requirement to protect and ensure security of plain text. Successful digital watermarking techniques are applied for images, speech, audio and video. In the same way, digital watermarking technologies are highly desirable for the plain text in order to uniquely verify the ownership.

Text watermarking also ensures protection for the content of digital text from illegal copying, copyright violations, infringements, redistributions and other similar tampering. Unlike the watermarking of image, audio, speech, video, text watermarking has limitation over capacity, since plain text contains less amount of redundant

information than other data. In order to provide sufficient confidentiality and robustness towards copyright protection and tamper proofing purposes, text watermarking techniques must satisfy the uniqueness and invisibility requirements [65].

3.6 Types of Text watermarking techniques

It is a great challenge to provide a simple mode of communication and exchange of information of a text to have copyright protection. When a text is transformed to another kind, the meaning, grammaticality, fluency, writing style and value of the text should not be changed. It is tough to protect small documents due to low capacity for watermark embedding. Therefore, the text watermarking algorithms highly depends on text size, rules, grammar, its language, writing styles and conventions. There are several techniques for text watermarking have been recommended including image based, semantic based, syntactic structure based, natural language processing based, word and sentence based, noun-verb based, zero watermarking algorithm, etc. The following sections describe the work done in the above mentioned category.

3.6.1 Text Image Based Watermarking

In the field of digital text watermarking, text document image is utilized for embedding the watermark. Text watermarking is difficult due to its sensitiveness, simplicity, and low capacity for watermark embedding. In the beginning, text document to be watermarked was treated as an image. Presently, researchers mainly rely on word-shifting and line-shifting techniques especially modification on spacing of letters, spacing of words, shifting of baselines modifying the kerns, serifs etc.,

In line shifting algorithm, the document image altered by moving lines downward or upward (left or right) depending on binary signal (watermark) to be inserted. This type of watermark can be extracted by using no-blind mode. In word-shift coding algorithm, the words within text are moved horizontal direction, which automatically expands the spaces to embed the watermark. Both the non-blind and blind modes are

suitable for this algorithm. And in the feature coding algorithm, watermark bits are encoded in order to alter certain text features in the text.

3.6.2 Syntactic Approach

Generally a text document is made up of characters, words, and sentences. Sentences have been formed in different syntactic structures. One of the approaches for text watermarking is the application of syntactic transformations on text structure to embed watermark, which is used in the past. In this approach, a syntactic tree is built initially then, some transformations are applied on this tree and through this the watermark is embedded in to the text without affecting any inherent properties of the text. It is also necessary to analyse the syntax and semantic structure of the text while changing the incorporated watermark bits, which are carried out by the natural language processing algorithms.

3.6.3 Semantic Approach

This approach uses the semantic structure of text for embedding the watermark. Here the text contents like nouns, verbs, prepositions, acronyms, words spelling, sentence structure, grammar rules, etc. are exploited during the insertion of watermark. They are language dependent and have limited applicability. Noun-verb based technique for text watermarking uses a grammar parser using semantic networks, which parses the nouns and verbs from a sentence.

3.6.4 Structural Approach

Structural approach utilizes the text constituents, their frequency, the ordering, and location to generate a watermark, hence the text is not being modified. For this purpose, zero watermarking has been used which constructs the watermarked text using the characters of original text without modifying the original text. Basically, language is composed of phrases, clauses, and sentences which are consisted of various types of words including pronoun, noun, verb, adjective, adverb, conjunction, preposition, and acronyms. Even, each word is constructed from alphabet, notation,

and symbols. Structural approaches utilize syntax and morphology for text watermarking. It is developed by using double letter. Firstly, the occurrence of each double letter ie., from AA to ZZ is counted in each group and the maximum occurring double letter is identified in each group. A list of maximum occurring letter (MOL) is formed that contains maximum occurring letter of each group with corresponding group number. Then, partial key is generated from 3-digit group size, 2-digit cipher attributes, and 1-digit cipher choice. The number of sentences to be included in one group is used for 3-digit group size. The watermark extraction is reverse process for embedding and encryption.

3.6.5 Hybrid Approach

Hybrid approach is developed to rectify the weakness of each single text watermarking approach by combining different text watermarking approaches. Hybrid text watermarking approach improves the robustness of text watermarking and also supports to watermark the wide text documents. Structural- and image-based approaches have been combined for this purpose in a specialized manner. The reason for such combination is due to similarity among pure alphabetical text watermarking [66].

3.6.6 Linguistic based approach

The linguistic-based approach is a natural language-based method, in which the syntactic and semantic nature of the cover text is changed in order to embed the watermark. The structure and meaning of the text remain unchanged in this watermarking. Linguistic-based watermarking mostly makes use of the semantic or syntactic transformation or a combination of both, depending on the language of the text.

In syntactic approach, the words in the set are manipulated to hide data. In order to hide the watermark message, the nouns, verbs, adjectives, synonyms, pronouns, prepositions, and other grammatical features of the text content are utilized. These alterations do not affect the original meaning of the text and the order of the words in

the sentences can be rearranged to hide bits. This approach embeds the watermark by altering the text structure, such as moving the adverbial phrase, adding the subject or changing the sentence from active into passive clause.

Semantic-based approach hides the data by manipulating the words in the text and embeds it as watermark. In this method like synonym substitution, algorithms based on typos, algorithms based on noun-verbs, acronyms and abbreviations, algorithms based on linguistic approaches of presuppositions, and algorithms based on text-meaning representational strings. This approach depends on properties of language used its vocabulary, grammar or structure to hide the watermark [67].

3.7 Biometric cryptosystem

Biometric cryptosystems (BCSs) are introduced to bind a secure digital key to a biometric which is generated digitally from biometric information. This scheme offers solutions for biometric-dependent key-release and biometric template protection for all kind of media. BCSs have replaced the password-based key release with several significant security benefits. It is more arduous to counterfeit, like copy, share, distribute etc., in BCSs. A user-group can be provided with equal level of security by all kinds of biometric characteristics. A number of “fuzzy comparisons” is being utilized in BCs by applying decision thresholds due to biometric variance. The decision thresholds are set up based on the score distributions between non-genuine and genuine subjects. Especially, BCSs are designed to provide 100% at authentication by means of stable output keys. In order to assist the key release process, the original biometric templates are replaced through biometric dependent public information.

The most BCSs require a helper data which are stored as biometric dependent public information in order to generate and retrieve key. It is not feasible to extract keys directly due to variance in biometric. Also, helper data does not reveal significant information about original biometric templates, during reconstruction of keys. An additional feature of BCs is authentication through verification of key validities, where the output is either a key or a failure message. The verification of keys means that biometric comparison which is performed in encrypted domain. Therefore, BCSs are

applied for biometric template protection and also to provide biometric-dependent key release. According to the generation of helper data, BCSs are classified in to two; key-binding or key-generation systems.

Key-binding schemes: Helper data are obtained by binding a chosen key to a biometric template. This binding process fuse the secret key and the biometric template, finally stored as helper data. At the time of authentication, these keys are extracted from the helper data by using a suitable key retrieval algorithm. A new helper data is generated by updating an existing key which requires re-enrolment, since cryptographic keys are independent of biometric features

Key-generation schemes: Helper data are derived from the biometric template. Here a helper data and a given biometric sample are used to generate a key. The major key-generation schemes store the helper data, in the case where helper data storage is not obligatory. If a key is extracted without the use of any helper data, such keys cannot be updatable.

3.7.1 Error-Correcting Output Codes (ECOC)

ECOC is a suitable tool to solve multi-class issues, which is used with multiple binary classifiers. Typical ECOC constitutes two phases: ECOC encoding, and ECOC decoding. Encoding phase assigns the distinct code-words to various classes, which forms an encoding matrix with dimension of $N_c \times n$, where n is the code-word length and N_c indicates the number of classes. In this matrix, each row indicates a class code-word and each column represents a dichotomizer.

Binary and ternary are the existing two popular designs which used for encoding and decoding. Binary encoding design consider the values of −1 or 1 as the elements for the encoding matrix whereas ternary encoding takes the values −1, 0 or 1 for to represent the encoding matrix. In both the designs, the ECOC matrix is encoded by +1 or −1 according to the class set membership. And in ternary design, if the class is not assigned by the dichotomizer, it will assign the encoded output as 0. In the decoding

phase, it first generates a code-word in range of $\{-1,1\}_{1\times n}$ for each test feature by applying n training binary classifiers. Then it compares this test code-word with the base code-word of each class which is in the ECOC matrix, and finally the closest test data is assigned as the required code-word.

3.7.2 Template protection method

3.7.2.1 Discriminant binarization transformation (DBT)

In face cryptosystems, the DBT plays an important role. It is designed with two major objectives. First one is to enhance the discriminability and the second one is to map a real valued template to its binary version. In this transform, initially an optimized ECOC matrix is generated using Dense Random encoding and an extension process. A target binary template for a subject is being described by a specified row of this matrix.

The Dense Random encoding initializes the code-word length of the ECOC matrix $M_{Nc\times n}$, and an iterative extension algorithm is utilized to increase it. A number of ECOC matrices are randomly generated in Dense Random encoding which in the range of $\{0,1\}_{Nc\times n}$, where N_C is the number of classes and $n = 10\times \log_2^{N_c}$. An initial encoding matrix is chosen which is based on highest minimum hamming distance between the matrix $M_{Nc\times n}$ and its code-word pairs. Then an extension matrix E is constructed and combined it with M iteratively, in order to extend the encoding matrix M. If E_t is the extension matrix obtained by iteration t, then the encoding matrix M at iteration $t+1$ will be $M_{t+1} = M_t \cup E_t$. A confusion matrix is obtained from encoding and decoding algorithms (Algorithms 1 and 2), and based on this matrix the extension matrix E which is obtained from each iteration is encoded. Finally, extension matrix E is constructed and encoded using the following steps

Step 1) Finding confused classes: First, determine the pair of classes with maximum confusion $\{C_i, C_j\}$ from confusion matrix; then extract $\{C_m | m = 1, \cdots, N_c \text{ and } m \neq i, j\}$ those who are confused with either C_i or C_j among all the other classes

Step 2) Calculating the number of extended columns k: Find $k = \log_2^p$, where $p = (\text{the number of confused classes with either } C_i \text{ or } C_j) + 1$.

Step 3) Associating code-words: Generate 2^k number of random binary code-words with length k; if two code-words $code_i$ and $code_j$ exhibits the maximum hamming distance, then $code_i$ and $code_j$, are associated to the confusion classes C_i and C_j. Finally, for other classes, random iterative code-words with length k, considering the overall confusion with all classes, $C_q = \{Cq' | q' = 1, \ldots, N_c, q' \neq i, j\}$, are associated.

Step 4) Satisfying ECOC conditions: the algorithm checks the equality of each column of $E_{Nc \times n}$ with $M_{Nc \times n}$ columns, after the extended matrix $E_{Nc \times k}$ is encoded (Step 3). In both matrices, if the columns are equal, then the algorithm exchanges the code-words of classes in $E_{Nc \times k}$ except those representing C_i, C_j and also those of classes which are confused with C_i or C_j.

The algorithm 1 depicted in figure 3.4 will be continued until one of the following conditions is satisfied: (i) length of the code-word reaches n, where $n = 15 \times \log_2^{N_c}$, or (ii) $C_{ij} = 0, \forall i, j$, meaning that all confusions between any two classes have been removed. Figure 3.5 shows the ECOC decoding process in detail.

Algorithm1: ECOC encoding and calculation of discriminant functions.

Input: Training data, ECOC matrix

Output: Discriminant functions

For to do *J=1 to n do*

1. Change training labels into two supper-classes using binary element with respect to each class in column *J of* $M_{N_C \times n}$
2. Compute the discriminant function, *Fj* , by training a linear SVM classifier with the relabelled training data (obtained in Step 1) using the following equation:

$$F_j(x) = W_j^T x + a_j \tag{3.6}$$

Where, is training template *Wj* is a weight vector in function *Fj* with the same length of sample *x* and aj is the threshold.

End for

Fig 3.4 ECOC Encoding Algorithm

Algorithm 2: ECOC decoding and calculation of confusion matrix.

Input: Labelled test data, ECOC matrix, $M_{Nc \times n}$,

Discriminant functions $F_1, F_2, \ldots, F_n$

Output: Confusion matrix

For $q = 1$ to the number of test data do

Extract the test binary template w_q by applying n discriminant functions $F_1, F_2, \ldots, F_n$

to the original test template, F_q, as follows:

$$F_j(f_q) = \begin{cases} 1 & W_j^T f_q + a_j > 0, j = 1, \ldots, n \\ 0 & W_j^T f_q + a_j \leq 0 \end{cases} \quad (3.7)$$

Where, W_j and a_j are weight vector and threshold in function F_j, respectively. The test binary template w_q with length n is formed by concatenating all of the bits calculated in Equation (3.7).

Compute the hamming distance between $w_i (i = 1, \ldots, N_c)$, being rows of $M_{Nc \times n}$, and binary vector w_q (calculated in Step 1). Hamming-distance

$$(w_i, w_j) = \sum_{j=1}^{n} sign(w_i(j) - w_q(j)) \quad (3.8)$$

Predict the class label of binary vector w_q, which will be the index of the vector w_i with the minimum hamming distance with vector w_q (obtained in Equation (3.8)).

End For

Compute the confusion matrix of the test data using the input class labels and predicted class labels.

Fig 3.5 ECOC Decoding Algorithm

3.7.2.2 Chaos feature permutation

The input of this method is binary template along with a user-specific key. First, the initial conditions of the chaos map are extracted from the user input key during the process of key generation. Then random sequence is generated to shuffle a binary

class code-word based on the index ordering of this sequence using the proposed chaos function and the extracted initial conditions [69]. Fig 3.4 explains the scheme of the chaos feature permutation method.

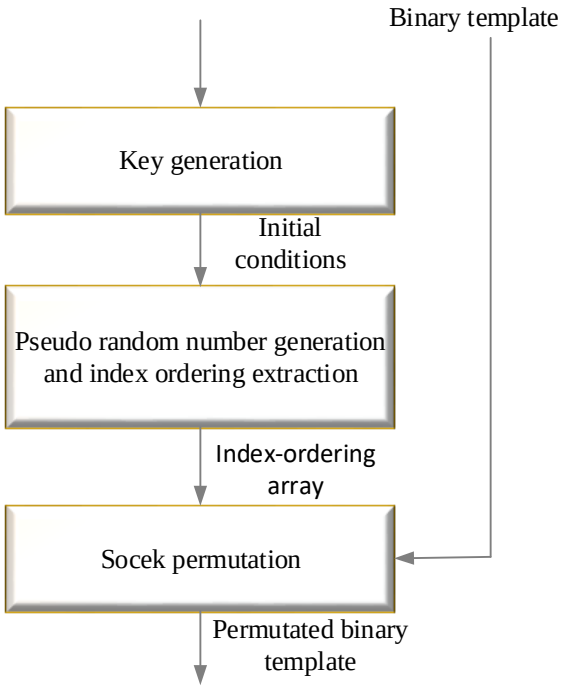

Fig 3.6 schema of the chaos feature permutation method

3.7.2.3 Key generation

A user key with 96 bits length has been employed in the process of key generation given as $K = K_1, K_2, K_3$, here each K_i constitutes 32 bits, hence three blocks are generated, each of which has 32 bits given by Q_i (32 bits) by the application of XOR operation on $K_{ij}\left(i = 1,\ldots,3 and\ j = 1,\ldots,4\right)$. These blocks act as initial conditions for the generation of the new chaos function, $\left(x_0, y_0, a_0\right)$ as follows;

$$x_0 = bi2de(Q_1)/2^{33} \tag{3.9}$$

$$y_0 = bi2de(Q_2)/2^{32} \tag{3.10}$$

$$a_0 = bi2de(Q_3)/2^{33} \tag{3.11}$$

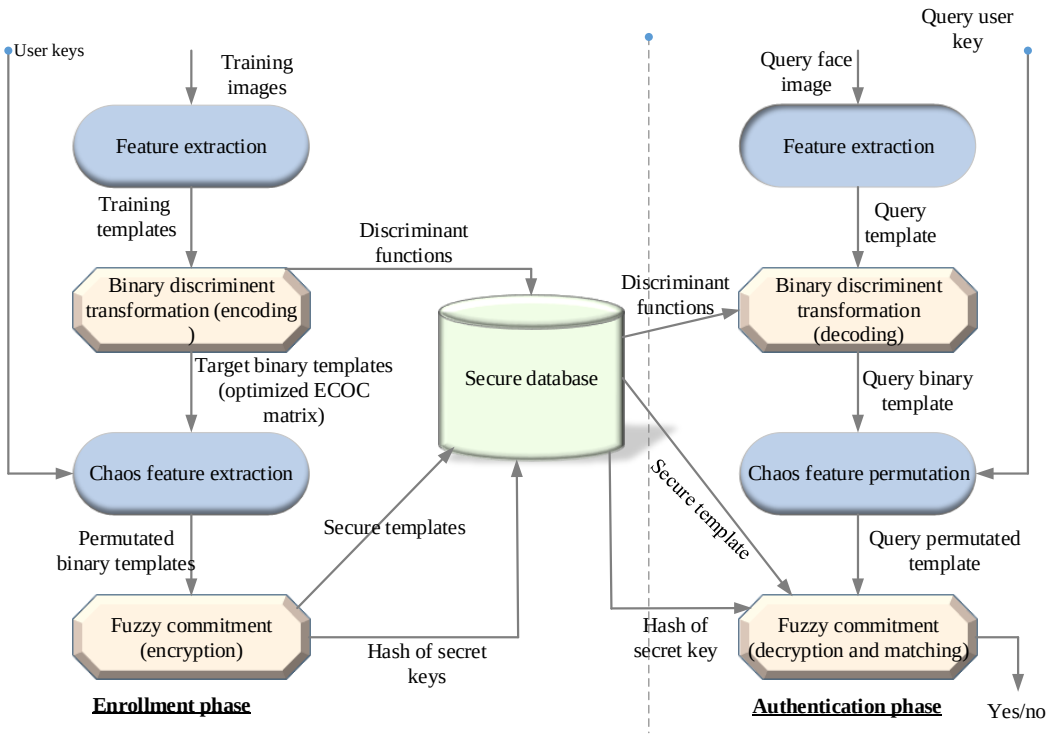

Fig 3.7 Overall structure of the biometric cryptosystem

The biometric cryptosystem has two phases: Enrolments phase and authentication phase. Figure 3.5 illustrates the overall structure of proposed biometric cryptosystem. In enrolment phase, encoding and encryption is processed. In authentication phase, binary discriminant transformation (decoding) fuzzy commitment (decryption and matching) has performed.

Algorithm 3: Enrollment phase in the biometric cryptosystem.

Input: training images, user leys key $(i = 1,2\ldots,N_c)$

Output: discriminant functions $(F_1,F_2,\ldots,F_n)$, secure templates $(P_i \oplus C_i, hash(C_i))$

Compute:

Extract the training templates, $fig(i = 1,2,\ldots N_c; q = 1,2,\ldots,q')$ from N_c classes where each class has q' samples.

Compute the optimized ECOC matrix $M_{Nc\times n} = [w_1, w_2,\ldots, w_{Nc}]^T$, including target binary templates, using ECOC extension algorithm

Call the ECOC encoding using training templates f_{iq} and $M_{Nc\times n}$ as input to compute discriminant functions $(F_1,F_2,\ldots,F_n)$.

For each target binary template $w_i(i = 1,\ldots,N_c)$ in $M_{Nc\times n}$ do

1. Apply chaos feature permutation with user key key_i on w_i to obtain the permutated template P_i.

2. Generate a random binary string C_i with the same length as P_i.

3. Encrypt P_i to $(P_i \oplus C_j, hash(C_i))$ using FCS.

4. Store $(P_i \oplus C_j, hash(C_i))$ in the database.

End for

Fig 3.8 Enrolment Algorithm for biometric Cryptosystem

Algorithm 4: authentication phase in the biometric cryptosystem

Input: query face image, discriminant functions $(F_1, F_2, \ldots, F_n)$, query key key_q , secure database

Output: authentication result

Compute:

Apply feature extraction process on a query face image to extract a test template f_q.

Convert f_q to its binary representation w_q, using equation 3 in algorithm 2 and the discriminant functions $(F_1, F_2, \ldots, F_n)$.

Permute w_q by using chaos feature permutation and its user key key_q, to obtain the permutated binary template P_q.

Release the corresponding reference data of query $(P_i \oplus C_i, hash(C_i))$, from secure database

Calculate $P_q \oplus (P_i \oplus C_i)$.

If $(hash(P_q \oplus (P_i \oplus C_i)) == hash(C_i))$

The authentication process succeeds

End if

Fig 3.9 Authentication Algorithm for biometric Cryptosystem

3.8 Possible attacks on various cancellable biometrics approaches

The main aim of the attacks is to obtain illegal access to some important data about the other user or original biometric feature. The major key point to ensure security of cancellable templates is the design of transformation function and the parameters associated with it.

The Stolen Token Scenario and Pre-Image Attack on Bio-Hashing: In biometric salting techniques, the original template can be reconstructed, if the user-specific key is compromised and the transformations are invertible. Bio-Hashing can be partially

solved by pseudo-inverse operation, because it is quantized under-determined linear equation system. A stolen token attack severely degrades the performance. If Bio Code and TRN are compromised, then a pre-image of original Bio Code can be generated by pre-image attack and inverse operations, even the biometric data belongs to not a genuine user. Reverse engineering methods allow invertibility, by which the inverses or pre-images can be easily calculated, leads to impersonation attacks.

The Stolen Token Scenario and Pre-Image Attack on Non-invertible Transforms:

Suppose the attacker availed with transformation key/parameter and transformation function, he can able to approximate the transformed template using the inverse. Then the attacker can generate a pre-image template equivalent to the original templates by performing several attacks on various non-invertible transforms.

Attack via Record Multiplicity (ARM): The attacker can retrieve some information from multiple transformed templates of the same biometric by correlation, even they can reconstruct the original template. A perfect knowledge on the stored templates along with transformation parameters and function determines the success of this attack. Recovering the original biometrics templates from the protected ones is computationally infeasible.

Privacy and Replay Attacks: Privacy attacks having great potential of threat to the privacy of users though it is passive. An attacker can locate a user through the determination of distorted biometric template which is used for a particular transaction. The adversary can predict whether it is the same person, but cannot find who is trying to be identified. Replay attacks are used to obtain illegitimate access, where the recorded signals are replayed in a link between sensor and matcher.

Dictionary Attacks: These attacks try to recover the original biometric template using two or more transformed templates. The attacker constructs a set of possible pre-images for each transformed template by simulating the transformation. Shin et al presented an algorithm on functional transformed based cancellable fingerprint templates for successful implementation of such attacks.

Hill Climbing Attacks: These attacks have designed to access systems communication channels and match scores. It inserts fake data into the system and modifies until the acceptable match score which is not generated by the system [70].

3.9 Conclusion

In the era of digital world, the computers and networks stores large amount of data when compared to hardcopy formats. Now a days the communication through the internet attained explosive growth. The soft copies of text documents stored in internet, sometimes affected by attacks such as illegal copying of important information, tampering and illegal authentication and redistribution of the copyrighted text documents. To rectify these types of attacks the researchers developed many techniques but among the all, watermarking technique provides solution for all type of threats. In this chapter we are providing the brief view of watermarking and security threats in data. This chapter explains the various stages during the process of watermarking and also the biometric cryptosystem utilized in watermarking technique.

CHAPTER 4
COMBINED MARKOV MODEL AND ZERO WATERMARKING FOR INTEGRITY VERIFICATION OF PDF ENGLISH TEXT DOCUMENTS

CHAPTER 4

COMBINED MARKOV MODEL AND ZERO WATERMARKING FOR INTEGRITY VERIFICATION OF PDF ENGLISH TEXT DOCUMENTS

4.1 INTRODUCTION

The fast growth of advance information technology has developed and made the digital access as simple. Due the growth the security of intellectual property rights of digital media has become an important problem. The watermarking method has attracted considerable attention and has lots of applications such as copyright protection, authentication, secret communication, and measurement [71]. The recent improvements in information and communication technologies gave numerous growths and simplified the digital content distribution, communication, and reproduction. However, some issues and threats are also occurred.

All digital multimedia contents in the Internet can be categorized into images, text, audio, and video, with the challenge being to assure protection and feasible communications for each media type [72]. The zero watermarking techniques achieving the authentication and verification of integrity for text files [73].

Most of the watermarking methods, no matter in spatial domain or frequency domain, change the original data during the watermark embedding process. For authority protection, some data is embedded in secret way but distorts the original data in the same moment. This case reflects a clash between invisibility and robustness. then we design the watermark in order to remove this conflict, based on the important properties of the text without data or original text changing. These types of approach called zero watermarking.

Recently, a new method of zero watermarking based on DWT and Markov model were developed [74]. Traditional encryption schemes do not have a full solution for

the issue of illegal copying because once encryption is deleted from a text file, there not capability for distribution control [75].

4.2 Zero Watermarking based techniques for integrity verification

4.2.1 Zero watermarking algorithms based on non-vowel ASCII characters

To watermark the text document, this algorithm utilizes the existence of non-vowel alphabets. Where source copyright owner of the text file generates a key based on watermark and structure of the content provided, utilizing an embedding algorithm. Since it generates the key for authors by utilizing characteristics of the text content without modification, it is referred as zero watermarking algorithms.

Initially, the text document is analysed and the corresponding prepositions are discovered. Then an average frequency preposition (AFP) is discovered followed by calculation of occurrence count. Later, the text partitioned of are created using AFP. The occurrence count of every alphabetical character is gained after partitioning and the maximum non-vowel ASCII characters are recognized and populate MONV list. Based on watermark provided by the owner, a specified key has been generated using MONV list.

In order to secure the copyrights of the owner, the generated author key is registered with a certification authority (CA). The original watermark and the author key are time stamped and fix aside with the CA. If anyone tries to tamper the copyrights of the author, this generated key is extracted and the original copyright owner is recognized, which can resolve the copyright clash. Sometimes one text may have more than one claim. This algorithm is suitable for both intelligent insertion and deletion attacks on the text.

To discover the owner or a group or an association who owns the copyrights of the text, the watermark utilized which generates a key that is to be pure

alphabetical and selects carefully. The original author performs watermark embedding and CA on the behalf of writers claim performs extraction.

Embedding Algorithm

The algorithm which embeds the watermark in to the text is known as embedding algorithm. Generally the watermark is applied logically in the text and finally based on AFP and non-vowel characters it generates the key.

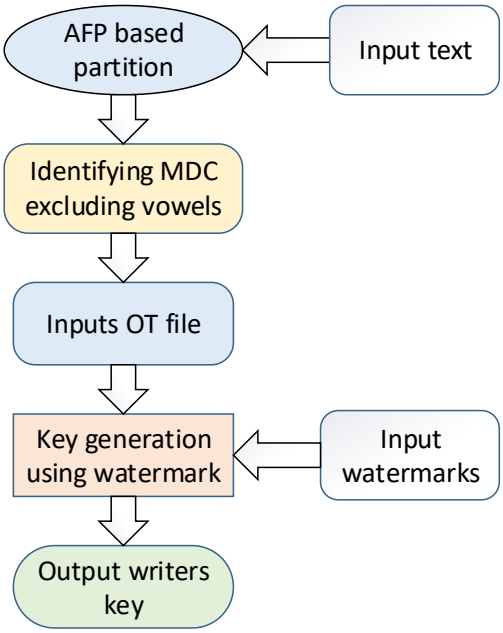

Fig 4.1.Embedding process zero watermarking

Extraction Algorithm

The extraction algorithm is utilized when the copyright clash occurs. From the noisy text this algorithm extracts the watermark, hence it is known as extraction algorithm. To find any type of copyright clash, CA uses the extraction algorithm. The author key utilized to extract the watermark from the tampered text by the help of this algorithm. Later, the original and the extracted watermarks are compared which results watermark accuracy and determines the original owner of the text document. The common strategy of watermark extraction process is described in fig 4.2. Initially, the author key is taken as the input, then the text is partitioned based on AFP which is

gained from author registered key. Later, the occurrence of non-vowel character in each partition is counted and a list is produced by which the highest occurring non vowel characters have been found. By using contents of author's key as described in the extraction algorithm, the watermark is obtained. Finally the copyright clash resolved by the help of CA [76].

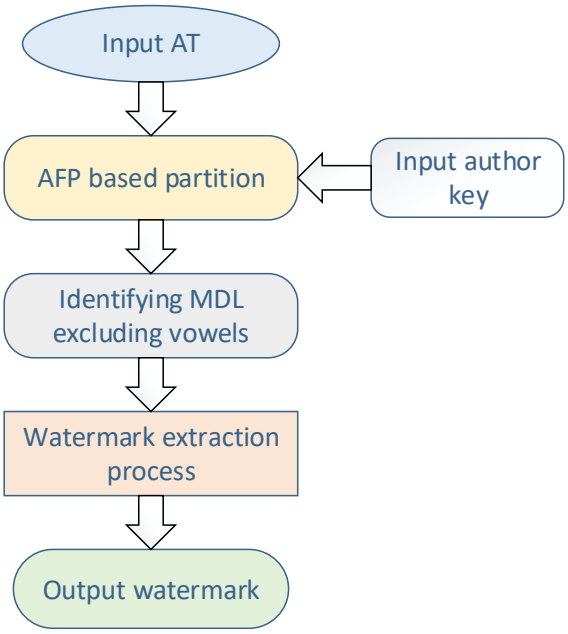

Fig 4.2.Extraction process zero watermarking

4.2.2 Zero-Watermarking algorithm on multiple occurrences of letters

The text watermarking methods targets to embed watermark information into the protective information such as text. For tamper detection and copyright protection, this information is utilized. A zero watermarking approach, which the host data is not modified to embed watermark, rather the characteristics of host text are utilized to create a watermark. The contents of text document are used to generate a watermark. Later, this watermark pattern matched with the pattern generated by tampered document to find any tampering. The watermark generation and extraction process is explained in fig 4.3. To detect tampering in the text document, the watermark is used in the extraction algorithm. The content of the text are utilized to secure text.

Attackers will always to try to tamper the document in such a way that the meaning remains same but the documents looks different. The attacker alters the text and change the place of various words such as nouns and verbs but cannot reject them to make a sentence.

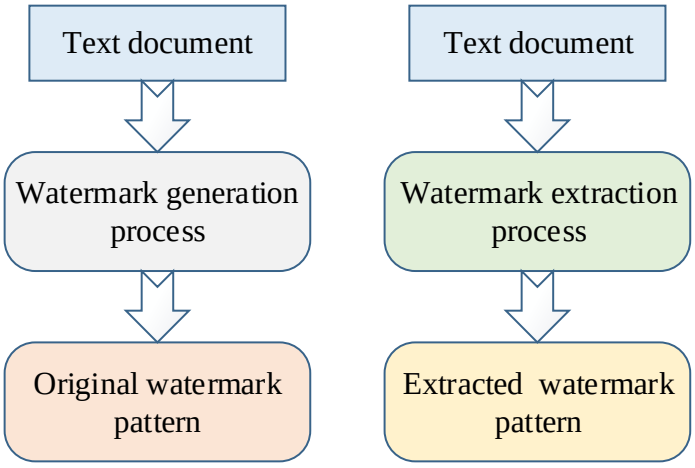

Fig 4.3.Embedding and extraction process

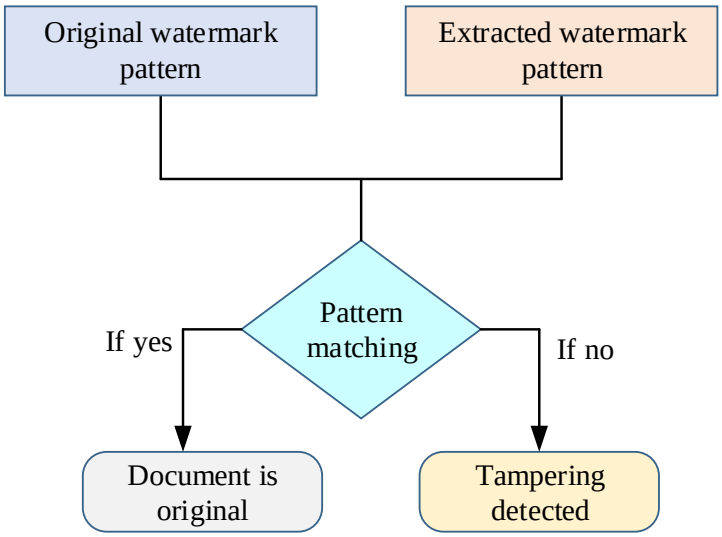

Fig 4.4.Pattern matching

In this algorithm, words with several occurrences of letters are found and the initial alphabets of those words are utilized to generate watermark patterns. Finally watermark is formed by concatenating these patterns.

Embedding Algorithm

The purpose of embedding algorithm is to embed the watermark in the text. In this algorithm the original text document is taken as input and watermark generated by this algorithm is known as output. The certifying authority plays an important role, which registers that watermark along with the original document, name of the author, date and time. Figure 4.5 describes the embedding algorithm in detail.

Embedding Algorithm

1. Read text file T1
2. NOS= Total number of sentences in T1
3. For $i = 1$ to NOS, repeat step 4 to 15
4. NOW=number of words in i th sentence
5. For $j = 1$ to NOW, repeat steps 8 to 15
6. LOW= length of j th word
7. For $k = 1$ to LOW repeat steps 6 to 15
8. SA=letter for comparison/search (k th letter)
9. Repeat-count=0;
10. For $m = 1$ to word-length, repeat steps 11 to 13
11. Currently-letter=m th letter of word
12. If(SA=current-letter)
    Repeat-count=repeat-count+1;
13. $m = m + 1$
14. if (repeat-count>1)
    word(k)=initial letter of k th word
    break internal for;
15. $k = k + 1$
16. $j = j + 1$
17. $i = i + 1$

output W(watermark)

Fig 4.5 Algorithm for Embedding Text Watermarking

Let us consider T1 be the original text document which is obtained from owner. In this document, some of the sentences are taken for which the length of each word is calculated. Then every single letter of a word is compared and evaluated separately with multiple occurrences of letters with words. The first letter of these words are combined to form watermark patterns. All these patterns are concatenated to generate the watermark. Finally generated watermark is registered to the certified authority.

Extraction Algorithm

1. Read text file T(A) and OW
2. Nos=total number of sentences in T(A)
3. for $i=1$ to NOS, repeat step 4 to 15
4. NW=Number of words in i th sentence
5. for $j=1$to NW, repeat step 6 to 15
6. LOW=length of j th word
7. For $k=1$to, LOW repeat steps 8 to15
8. SA-letter for comparison/search (k th letter)
9. Repeat-count=0;
10. For $m=1$ to word-length, repeat steps 11 to 13
11. Current-letter=m th letter of word
12. If(SA=current-letter)
    Repeat-count=repeat-count+1;
13. $m=m+1$
14. if(repeat-count>1)
    work(k)=initial letter of kth word
15. $k=k+1$
16. $j=j+1$
17. $i=i+1$
18. if EW=OW(primary match)
      PM(p)=1
    else
      if EW=OW (secondary match)
          PM(s)=no of matched secondary patterns/TP
      else

PR=(NM(P)+NM(s))/TP

Fig 4.6 Algorithm for Extracting Text Watermarking

T(A): Attacked text file, PM: Pattern matching rate, TP: Total patterns, OW: Original watermark, EW: Extracted watermark, NM: Number of matched patterns.

Extraction Algorithm

The extraction algorithm is intended for extracting the watermark from the text in an effective way. In the extraction algorithm the text document is taken as input and extracted watermark will be the output. This text document might have attacked by intruders. Initially a watermark pattern is generated from this text document. Later, this watermark pattern is compared with the patterns of original watermark [77].

The authentication of the document detected accurately by this algorithm which is described in figure 4.6. This method avoids the authors from common sentence re-writing and re-ordering attacks. However, it is also possible to destroy the watermark due to exceeding limit of tampering attacks.

4.3 Combined zero watermarking and Markov model

A novel algorithm which uses digital watermarking concept and Markov model has discussed in this section. This algorithm named as LNMZW3 and WNMZW4. In LNMZW3 algorithm, the original text document is not modified to embed watermark which means that a logical watermark embedding process is used. The LNMZW3 algorithm utilizes the Markov model for the natural language that is Markov chains which are utilized to check the contents of text documents in English. Then it extracts the probability features as probabilistic patterns of interrelationships between these contents, based on letter mechanism and third order of Markov model. The watermark key is generated using the probability features which is logically embedded with the original text document or stored in the watermark database. Any type of tampering that have happened in the document and unauthorized accessing of its content can be detected by comparing the embedded watermark key with the watermark key generated from attacked document. These steps are described in fig 4.7.

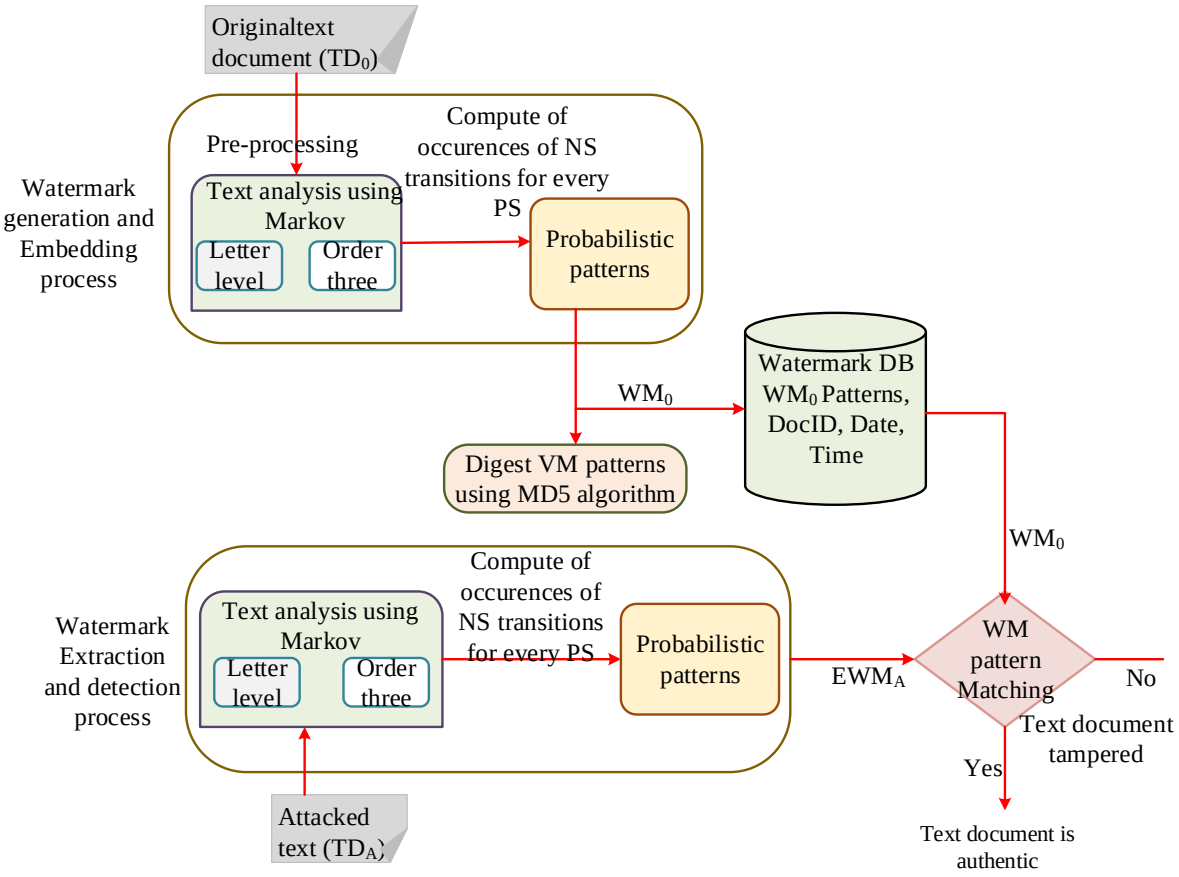

Fig 4.7 Markov model based Watermark generation and detection process

This section describes method of modelling English text using our LNMZW and WNMZW algorithm with Ngram order of Markov model. Let us consider an example of an English sentence as shown in Figure 4.8.

"The quick brown fox jumps over the brown fox who is slow jumps over the brown fox who is dead."

Fig 4.8 sample text

Each sequence of three unique consecutive letters in the given text is considered as a state by itself, when we use 3-gram of LNMZW. The process of Markov transitions from a state to another state as the text is read.

When we proceed the 3-gram order of Markov model, LNMZW3 for the above sample text, the list of all possible states with unique letter-level of 3-gram order obtained as follows.

"the", "he ", "e q", " qu", "qui", "uic", "ick", "ck", "kb", "kbr", "bro", "row", "own", "wnf", "info",..., "dea"

If the Markov chain is currently at "the q" state, then the possible transitions for each a state based on the above process are,

"the →, the → q, the → u, the → i, the → c, the → k, the →, the → b, the → r, the → o, the → w, the → n, the → d"

The existing scenario of our LNMZW3 algorithm is utilized to construct all possible states and their transitions can be described as Markov chain matrix $M[i][j]$. For the above example of sample English text contains 94 characters with spaces.

If we use various Ngram orders of Markov model based on word mechanism, each single word refers to 1-gram or one order of Markov model. The length of the Ngram refers to the number of word sequence in each unit of the text represented as an individual state within Markov chain. When we use 1-gram, each unique word in the given text is a state by itself, and when we use 2-gram, each sequence of two unique consecutive words in the given text is a state by itself.

Similarly, when we use 3-gram order and 4-gram order, each sequence of three unique consecutive words and four unique consecutive words is a state.

Also, when we use 4-gram of Markov model to process the same sample text presented above, the system produces a list of all unique of three consecutive word sets as possible states as the following:

"the quick brown fox", "quick brown fox jumps", "brown fox jumps over", "fox jumps over the", "jumps over the brown", "over the brown fox", "the brown fox who", " brown fox who is", "fox who is slow", ...,"fox who is dead."

When the above sample text is processed and if the Markov chain is currently at “the quick brown fox” state, then the possible transitions for each state that could come next are;

“the quick brown fox→jumps, the quick brown fox→over, the quick brown fox→the, the quick brown fox→brown, ..., the quick brown fox→dead.”

Thus, the same cases when we use 5-gram or more of Markov model to analyze English text documents.

“the quick brown fox jumps”, “quick brown fox jumps over”, “brown fox jumps over the”, fox jumps over the brown”, “jumps over the brown fox”, “over the brown fox who”, “the brown fox who is “, “brown fox who is slow”, “fox who is slow jumps”, ..., “brown fox who is dead”

As far as this sample text is processed, and if the Markov chain is currently at “the quick brown fox” state, the possible transitions for each state that could come next are;

“the quick brown fox jumps→over”, “the quick brown fox jumps→the”, “the quick brown fox jumps→brown”, “the quick brown fox jumps→fox”, ...,“the quick brown fox jumps→dead”

When the system completes the process and the corresponding Markov chains are represented by equating the expressions (1), (2), (3) and (4), 49 unique states U_s can be found. Also 21600 possible states P_s as sets of three consecutive letters, 61 possible transitions for each state T_s and 2989 possible transitions for all states P_t are obtained.

After being processed by the system, and represented in the Markov chains by equating the expressions (1), (2), (3) and (4), we get for 49 unique states U_s and we can obtain for 21600 possible states P_s as sets of three consecutive letters, 61 possible transitions for each state T_s, and 2989 possible transitions for all states P_t.

In the case of WNMZW3, we get for 12 unique states U_s, we get 6800 possible states P_s as sets of three consecutive words, 20 possible transitions for each state and 136000 possible transitions for all states P_t.

In the case of WNMZW4, we get for 12 unique states U_s, we get 128000 possible states P_s as sets of three consecutive words, 20 possible transitions for each state and 2560000 possible transitions for all states P_t.

$$P_s = (n-3)^3 \tag{1}$$

$$U_t = (n - R_s - 3)^3 \tag{2}$$

$$T_s = \sum_{i=1}^{i=n} U_s(i) + 3 + 3(3-1) \tag{3}$$

$$P_T = P_s * T_s \tag{4}$$

Where,

- n - total count of all character sets in the given text document.
- P_s - total of all possible states.
- U_s - total unique states (letter sets) within the given text document.
- R_s - count of all the repeated states within Markov chain matrix.
- T_s - all possible transitions for each state.
- P_T - total possible transitions for all states.

Based on the 3-gram order of Markov model on letter level to check the sentence, table 4.1 obtained. After analysing the above given sentence, from the table 4.1 if we consider a state "*jum*", the possible next state transitions are “p”, “p”, which means that the transition “p” occurs twice. Next a simple method is used to build the states and the Markov transition matrix $m[i, j]$. This is the most important part of text analysis using Markov model.

Table 4.1.Sample text states and transitions of LNMZW3

State ID	state	Transitions
1	(“br”):	([‘0’, ‘0’, ‘0’])
2	(“bro”):	([‘W’, ‘W’, ‘W’])
3	(“dea”):	[‘d’],
4	(“ick”):	[‘’],
5	(“jum”):	[‘p’, ‘p’],
…	…	…
…	…	…
…	…	…
48	(“xj”):	[‘u’]
49	(“xw”):	[‘h’, ‘h’],

As a result of analysing the above given sentence with the Ngram orders of Markov model based on word level, these table 4.1 and 4.2 are obtained. Table 4.1 describes the results of analysing the above mentioned sentence based on the WNMZW approach with fourth order and fifth order Markov model named as WNMZW4 and WNMZW5.

Table 4.2: Sample text states and transitions of WNMZW4

State ID	State	transitions
1	(“the quick brown fox”)	[“jumps”]
2	(“quick brown fox jumps”)	[“over”]
3	(“brown fox jumps over”)	[“the”]
4	(“fox jumps over the”)	[“brown”]
5	(“jumps over the brown”)	[“fox”, “fox”]
6	(“over the brown fox”)	[“who”, “who”]
7	(“the brown fox who”)	[“is”, “is”]
8	(“brown fox who is”)	[“dead”, “slow”]
9	(“fox who is slow”)	[“jumps”]
10	(“who is slow jumps”)	[“over”]
11	(“is slow jumps over”)	[“the”]
12	(“slow jumps over the”)	[“brown”]

From the above table 4.2 if we consider the state “brown fox who is”, the next state transitions are “dead” and “slow”.

Table 4.3: Sample states and transitions of WNMZW5

State ID	State	transitions
1	(“the quick brown fox jumps”)	[“over”]
2	(“quick brown fox jumps over”)	[“the”]
3	(“brown fox jumps over the”)	[“brown”]
4	(“fox jumps over the brown”)	[“fox”]
5	(“jumps over the brown fox”)	[“who”, “who”]
6	(“over the brown fox who”)	[“is”, “is”]
7	(“the brown fox who is”)	[“dead”, “slow”]
8	(“brown fox who is slow”)	[“jumps”]
9	(“fox who is slow jumps”)	[“over”]
10	(“who is slow jumps over”)	[“the”]
11	(“is slow jumps over the”)	[“brown”]
12	(“slow jumps over the brown”)	[“fox”]

From the above table 4.3, if we consider the state “the brown fox who is”, the next state transitions are “dead” and “slow”.

4.3.1 Watermark patterns generation

Generally the watermark generation and embedding algorithms takes the original text document as input. Then all capital letters are converted to small letters and all the spaces are removed from the original text document in the pre-processing step. The watermark is the output of the algorithm. A watermark database is being utilized to

store the generated watermark along with the original text document, document identity, author name, current date and time.

The process of watermarking constitutes two steps: watermark generation and watermark embedding. Typically a watermark is generated from the original text document and embedded logically into original text using embedding algorithm. In this proposed algorithm, the original text document (T) is provided by the author. Markov model is used for the text analysis process which will compute the number of occurrences of the next state transitions (ns) for every present state (ps). All the transition probabilities are represented in a matrix which contain number of occurrences of transition from a state to another computed by equation (5)

$$M[ps][ns] = P[i][j],\ for\ i,\ j = 1,2,\ldots,n \tag{5}$$

Where,

n - total number of states

i - refers to PS "the present state".

j - refers to NS "the next state".

$P[i,\ j]$- probability of making a transition from character set i to character j .

Once the text analysis and probability features extraction have completed, the watermark can be obtained by identifying all the non-zero values in the matrix. Then to generate a watermark patterns, the non-zero values are sequentially concatenated which is denoted by WMP_O as given by the equation (6) and presented in figure 4.9.

$$WMP_O \& = M[ps][ns],\ for\ i,\ j = non\ zero\ values\ in\ the\ markov\ matrix \tag{6}$$

Finally the sequential patterns of watermark as presented in figure 4.9, is stored in a watermark database along with the original text document, document identity, author name, current date and time.

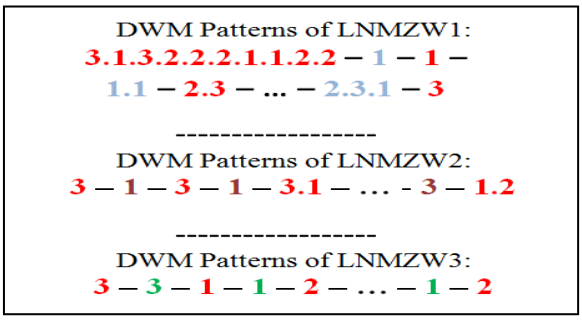

Fig 4.9 The original watermark patterns of the given text sample based on LNMZW

This procedure obtains all nonzero values of each state from the Markov chain matrix and concatenates them sequentially to generate the original watermark patterns based on WNMZW approach with 4gram and 5gram orders presented in figure 4.10

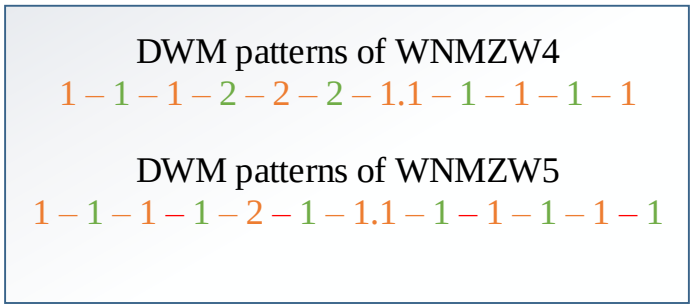

Fig 4.10 The original watermark patterns of the given text sample based on WNMZW4 and WNMZW5

After a sequential pattern of watermark has generated, in order to obtain secure and compact form of watermark, an MD5 message digest is generated given by the equation (7).

$$WMP = MD5(WMP_O) \tag{7}$$

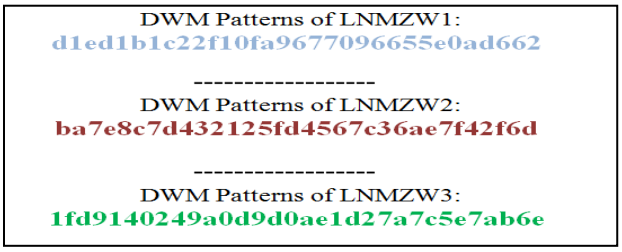

Fig 4.11 The original watermark patterns after MD5 digesting

Algorithm1: Watermark generation and embedding algorithm

Watermark generation

- Input: matrix of the transition probability from i^{th} word sets to j^{th} words $(MP[ps][ns])$

-output: WMP_O, Digests watermark WM_O.

-BEGIN

- // Analysis of the given text by computing the total frequencies of the transitions for every state

- Call text_analysis $(T_P, Ngram)$

- // concatenate watermark patterns of every states shown in Markov chain matrix

- Loop $cs = 1 \text{to } MP[states].Length$

  - Loop $ns = 1 \text{to } MP[transitions].Length$
    - IF $MP[cs][ns] != 0$ // states that have non zero transitions
    - $WMP_O \& = MP[cs][ns]$
  - $ns++$

- store WMP_O in DWM database

- // Digest the original watermark using MD5 algorithm

- $WM_O = MD5(WMP_O)$

- Output WMP_O, WM_O

- END

Where, WM_O refers the original watermark, WMP_O refers original watermark patterns and MD5 refers hash function.

4.3.2 Watermark Extraction and Detection Algorithm

The watermark detection algorithm must work towards zero-watermark. The proposed algorithm generates a watermark patterns 2 for attacked document for detection about the attacks. When watermark patterns 2 is received, which is compared to the watermark already stored with the text document. The matching rate of patterns 2 and the calculated watermark distortion can detect the tampering and content authentication.

Watermark extraction and detection can be done in two steps. First it extracts the watermark from the received attacked text document using watermark algorithm. Next it matches the extracted watermark with the original watermark using detection algorithm [78].

In the proposed watermark extraction algorithm also doing the same thing. First it takes the attacked text document, and finds a new watermark pattern using watermark generation algorithm. Then detection of watermark is performed.

Algorithm2: Watermark extraction

DWM Extraction
-Input: T_A, *Ngram order*
- Output: WMP_A
- BEGIN
-// perform pre-processing process for attacked text document
-call pre-processing (T_A)
- analysis of attacked text by computes the total frequencies of the transitions for

every state

- call text_analysis $(T_A, Ngram)$

- //generate the attacked watermark patterns from the attacked text document

- Loop $cs = 1 \text{to } MP[states].Length$

  - Loop $ns = 1 \text{to } MP'[transitions].Length$
    - IF $MP'[cs][ns] != 0$
    - $WMP_A \& = MP'[cs][ns]$
  - $ns++$

- $cs++$

- Output WMP_A

- END

WMP_A, refers attacked watermark patterns, T_A refers attacked text file, $MP'[cs][ns]$ indicates Markov matrix of the attacked text document.

Algorithm 3: Watermark detection

DWM detection

-input: pre-processed text $(T_P, T_P', Ngram)$

- output: PMR, WDR

- BEGIN

- //getting watermark patterns of the original document

-call watermark generation $(MP[cs][ns])$

- // extract watermark patterns from the attacked document

-call watermark extraction $(MP'[cs][ns])$

- // perform matching process between the original and attacked watermark patterns

  - IF $WM_A = WM_O$

    - Print "Document is authentic and no tampering occurred"
    - PMR=1

  - Else

    - Print "Document is not authentic and tampering occurred"

- // compute pattern matching rate (PMR) on the transition level

  - Loop $i = 1$ to $MP'[cs].Length$

    - Loop $i = 1$ to $MP'[ns].Length$
    - IF $WM_A = WM_O[i][j] != 0$

      - $patterncount += 1$

$$PMR_T(i, j) = \left| \frac{WMP_O[i][j] - (WMP_O[i][j] - WMP_A[i][j])}{WMP_O[i][j]} \right|$$

    - $transPMRTotal += PMR_T$

- Else

  - IF $WMP_A[i][j] != 0$
  - $patterncount += WMP_A[i][j]$

- // compute pattern matching rate on state level

- $$PMR_S(i) = \left| \frac{\sum_{j=1}^{n} (PMR_T(i, j))}{total\ statement\ \ pattern\ count(i)} \right|$$

- $stateweight(i) = \frac{PMR_S * Transitions\ frequency(i)}{total\ number\ of\ transitions}$

- $S_W += Stateweight(i)$

- // compute pattern matching rate on document level

- $PMR = \frac{\sum_{i=1}^{n}(S_W) * total\ number\ of transitions}{total\ number\ of\ transitions}$

- // compute watermark distortion rate on document level

- $WDR = 1 - PMR$

-END

Where, S_W : is the weight of states correctly matched. WDR represents the value of the watermark distortion rate.

4.4 Experimental setup and results

The accuracy of the tampering detection of this proposed system is being essential to calculate and compared with the existing algorithms i.e., previous orders of letter level and word level of Markov model. To compare the performance, authentication features, most general nature and volume of possible tampering attacks including insertion, deletion and reorder attacks are considered in several irregular places of the experimental datasets. This part describes the results and calculation of the LNMZW and WNMZW algorithms.

The performance evaluation of the LNMZW and WNMZW algorithm at various tampering attacks on different dataset sizes, were conducted for various situations in terms of type of attacks and volumes. Then, the average performance was determined and compared with previously existing algorithms which are LNMZW1, LNMZW2 and WNMZW4. Finally the better performance towards discovering the attacks is

found. The different kinds of attacks have been verified namely insertion, deletion, and reorder attacks in several random places in the experimental datasets [79].

Table 4.4 Dataset names and sizes attacked types and volumes

Dataset name	Dataset size	Attack types and volumes		
		Insertion	deletion	reorder
Chars 74k	47	5%,10%, 20%, 50%	5%,10%, 20%, 50%	5%,10%, 20%, 50%
ICDAR(full)	62			
ICDAR(50)	76			
ICDAR(CH)	52			
Street View Text	57			

The experimental datasets are namely Chars 74k, ICDAR (full), ICDAR (CH), ICDAR (50) and Street View Text (SVT). The datasets are a large collection of writings of a specific real time data. The datasets were verified for all scenarios of attacks commonly different volumes and types. Here, the datasets are used with 5%, 10%, 20%, and 50% of the attack volumes used in all kinds of attacks such as insertion, detection, and reorder attacks are determined. The datasets and attacks details used here are described in Table 4.4.

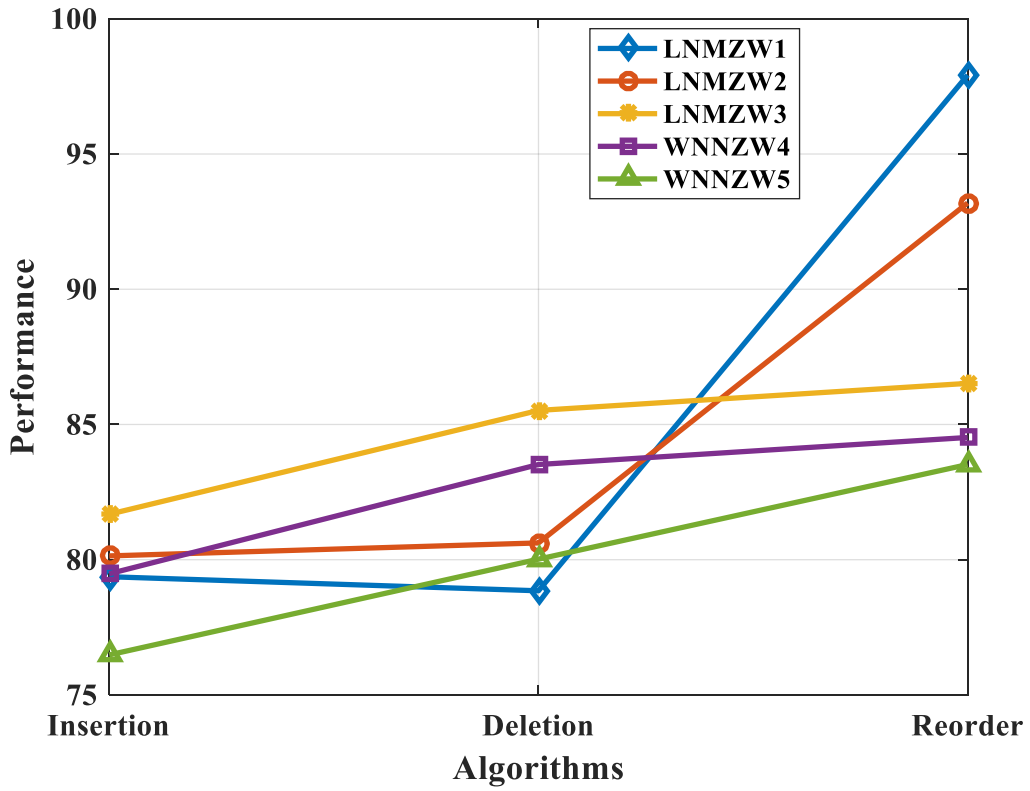

Fig 4.12 performance average of LNMZW1, LNMZ2, LNMZW3, WNMZW4 and WNMZW5 algorithms under all attacks

This figure 4.12 shows the performance averages of existing algorithms against dataset of different sizes and scenarios of reorder with all attack volumes. Figure 4.12 shows the general Markov model of LMNZW 1, 2, 3 and WNMZW4, 5 which are attained high performance under reorder attack. The performance of LNMZW1 algorithm is higher than LNMZW2, LNMZW3, WNMZW4 and WNMZW5 for reorder attack. In contrast, the LNMZW3 algorithm has higher performance than the algorithms LNMZW1 and LNMZW2, under insertion and deletion attacks.

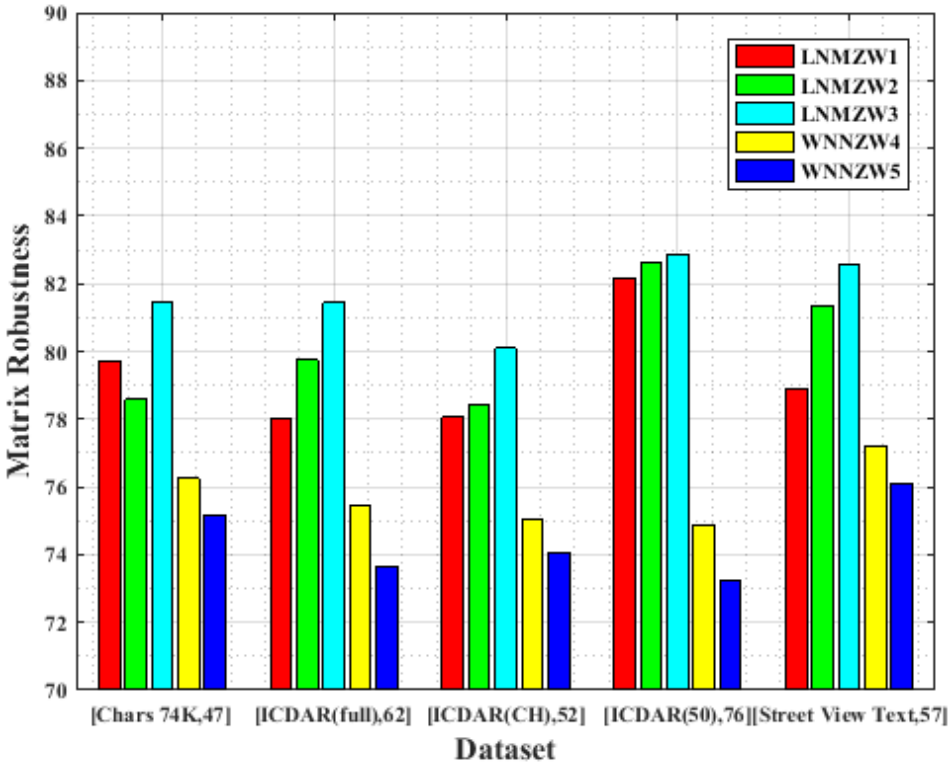

Fig 4.13 comparison of data set size effect on watermark robustness under all volumes of insertion attack

The figure 4.13 shows the comparison on the effect dataset size which determines the robustness of watermark for LNMZW and WNMZW algorithms under insertion attack. In figure 4.11, the dataset named [ICDAR (50), 76] has the higher robustness and has an average value 82.87% with LNMZW3 algorithm. The result of this graph shows that a slight impact is found due to the changes in the size of datasets. It also indicates that the documents which are large in size attains the watermark robustness more positive effects. This means that the watermark robustness value decreases when the document size decreases and vice versa. This graph shows that based on the datasets the robustness value of the compared algorithms will change.

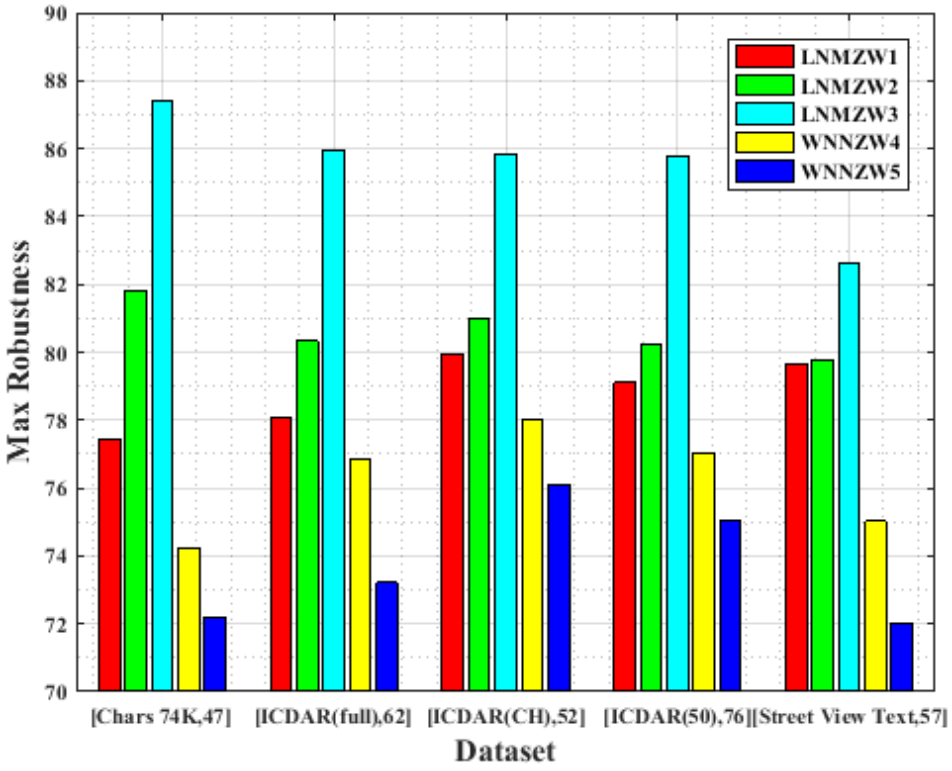

Fig 4.14 comparison of dataset size effect on watermark robustness under all volumes of deletion attack

This figure 4.14 shows the comparison of dataset size effect on watermark robustness under different volumes of deletion attack. From figure 4.14, we can observe that the smallest dataset named [Chars 74k, 47] has better watermark robustness with an average value 87.43%. The results show significant impact on the watermark robustness due to different volumes of deletion attack with different datasets sizes. In case of both LNMZW and WNMZW algorithms, small documents has the higher watermark robustness value and large documents has the lower robustness. While in the case of LNMZW1 algorithm, the large documents attains higher robustness value and small documents attains smaller robustness.

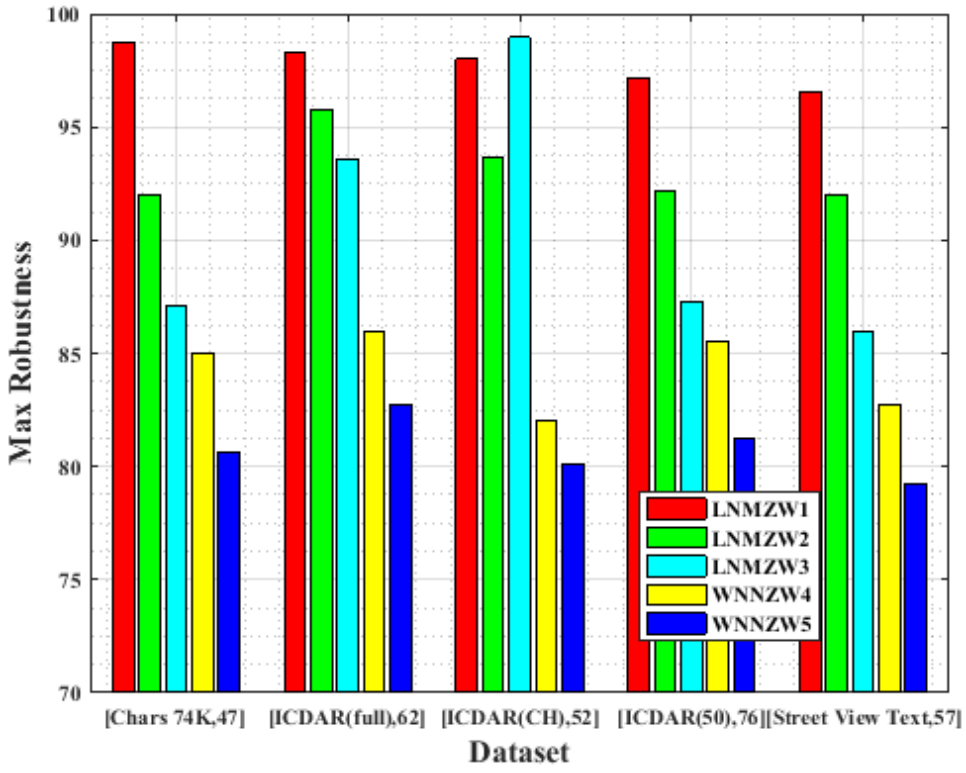

Fig 4.15 comparison of dataset size effect on watermark robustness under all volumes of reorder attack

Figure 4.15 depicts the comparison on the effect of dataset size in watermark robustness under different volumes of reorder attack. This graph indicates the average of maximum watermark robustness value for LNMZW and WNMZW algorithms and made comparison with previous algorithms against all dataset sizes and all scenario of reorder attack. Here a better average of watermark robustness value for all datasets is found with smaller documents. While a lower average watermark robustness values are obtained for larger documents for all the cases.

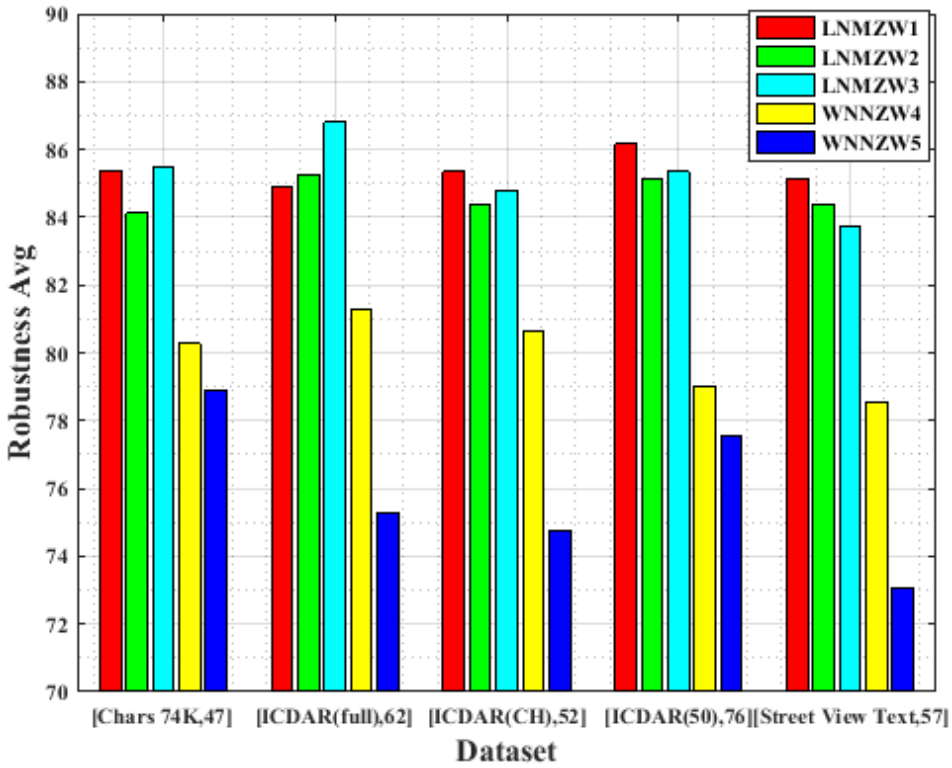

Fig 4.16 Comparison of dataset size effect on watermark robustness under all volumes of all attacks

In figure 4.16 the average watermark robustness values with different dataset size under all volumes of all attacks. The experimental results show that, when the Ngram order of Markov model decreased the robustness value improved and vice versa. Based on the observation, the impact on dataset size on watermark robustness in case of our WNMZW approach is improved when the size of the text document increases. The watermark robustness is improved when the size of text document decreases in the case of LNMZW approach.

4.5 Conclusion

In this chapter the Markov model based watermarking is described. Third order of LNMZW and fifth order of WNMZW is utilized to analyze the English text. First, the probabilistic patterns of states and transitions and corresponding inter-relationship of contents of the text documents are obtained. Finally the watermark is generated and

detected. In this study the LNMZW and WNMZW algorithms is compared with existing LNMZW1, LNMZW2 and WNMZW4 approaches against different datasets. The graph result shows that the reorder attack outperforms insertion and deletion attacks in terms of maximum average of watermark robustness with LNMZW and WNMZW approaches. The results show that WNMZW approach is suitable with small document sizes and LNMZW approach is suitable with all document size under different attacks.

CHAPTER 5
PROPOSED HSW BASED ZERO WATERMARKING

CHAPTER 5

PROPOSED HSW BASED ZERO WATERMARKING

5.1 Introduction

Currently, the information content is being stored and distributed in the form of digital information. Recently, highly developed internet services helps to handle various types of materials such as images, videos, e-books, e-journals and other digital contents. The demand for these content materials is going on increasing, hence these require copyright protection and integrity verification of all digital contents [80]. Digitally stored data is always under the risks such as can be copied and redistributed by some other name. Also digital data can be certainly modified and manipulated with the help of advanced technologies in computer technology.

Encryption is the way to secure the digital information, but once the data is decrypted it is very simple to copy and utilize the same data again [81]. Digital watermarking is another solution for this issue of copyrighting of data. The watermarking is a method of embedding information in any digital media such as text, image, audio or video. Based on that, the information can be easily replaced in web. To manage the copyright and to afford integrity to different types of digital data, the digital watermarking techniques offer various methods. The approved user can privilege to the originality of the file by applying any necessary watermarking procedure, if any attacker modified the data or duplicates it [82].

These digital watermarking methods are globally emerged to secure the multimedia copyright. In this digital watermark method, the embedding process inserts the watermark (copyright, signature or label) into different kinds of media. In this watermarking system, there are two methods were mainly used. Embedding of the watermark into the original data and extracting the watermark from watermarked data or attacked watermarked data. When constructing a watermarking scheme, some of the basic needs must be taken into accounts which vary depending on the use of the

scheme. The requirements of key watermarking are imperceptibility, robustness, security, and capacity.

The total number of hidden bits in an object is the capacity. Imperceptibility is utilized to calculate the dissimilarities between the original and watermarked object by noting any addition to the original object. After the watermark has been attacked, the robustness is the ability to extract or detect the watermark. In extraction process the security requirement without the destruction of the watermarked object is the difficult one. Due to the relative lack of unnecessary information within a text file compared to an image or audio files, a hidden watermark creation in the text is difficult. The human sensitivity to text changes is higher than the sensitivity to other multimedia. Any text change must reserve the meaning, fluently, writing style and text value [83].

The text watermarking techniques are classified into several methods. They are an image-based, syntactic, semantic and structural method. On image-based method, the watermark information is embedded with the text image. The syntactic arrangement of the textual content is used to create the watermark in the syntactic technique. Semantics of textual materials are used to create a watermark on the lexical method. Structural watermarking technique is a latest watermarking method that uses the composition of textual materials to create watermarks. After embedding this kind of watermark the watermarking method can never change any content and this technique is not valid on some textual content like web content, temporary point and mathematical notation with authorized text files. The content of the digital medium is modified with the traditional watermarking techniques by enabling watermarking approaches. But for practical scenarios it is not applicable for the plain text. The specific need for plain text is accomplished with the method of zero watermarking. This approach does not change the properties of original data but it uses the properties of original for generating watermark information. A novel technique HSW is produced in this chapter for tampering detection of plain text contents and it develops the contents of text itself for its authentication. The tampering attacks which include

deletion, insertion and reordering can be prevented and the material authenticity is attested with the help of suggested algorithm.

5.2 General watermarking

The author information is embedded as secured information by the help of watermarking technology. The document can protect without disturbing the meaning and appearance of the text from illegal use. To avoid this problem the information of the original author may be hidden in the text itself by natural language watermarking. Watermark robustness is improved through text meaning and the edit distance method. The basic method of watermarking involves embedding the secret key into the text so that original document protected from illegal use. The important thing that keeps watermark secretly in text which is a secret key utilized to embed watermark in text. Comparing with embedding the watermark extraction is quite complex.

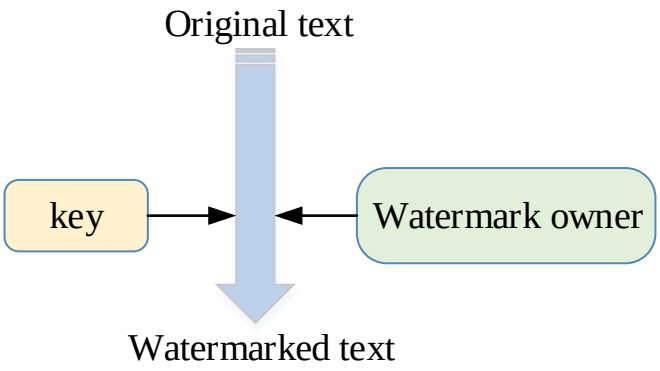

Fig 5.1 General watermarking method

Figure 5.1 describes the general watermarking technique. By using watermark key the original text is watermarked and converted into watermarked text.

5.3 Zero watermarking algorithm based on structural component

Recently structural component based approach is the utilized text documents watermarking. This approach includes encryption, steganography and watermarking. Since this approach uses of alphabetical watermark, it is not useful for all types of text documents. This technique includes two processes.

5.3.1 Watermark embedding process

In this process the watermark image and text file is given input to the image to text converter. The output of the converter (watermark alphabet) will give input of the grouping process. The watermark alphabet, preposition and group size will generate MOFL list. Based on that watermark key is generated as shown in figure 5.2.

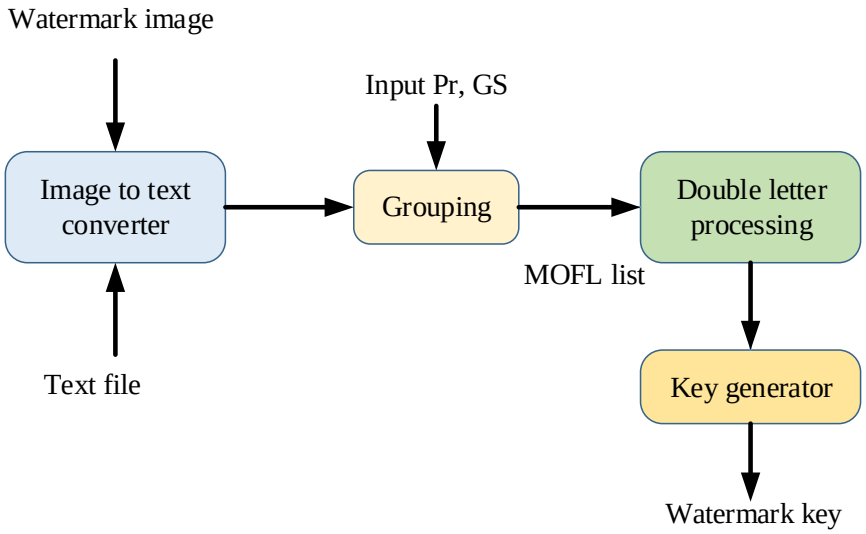

Fig 5.2 General watermark embedding process

5.3.2 Watermark extraction process

In this extraction process watermark key and the text file will be given as an input for grouping. Then the MOFL list is created than watermark generator generates the watermark alphabet [84]. Then the text to image converter converts the text into watermark image. The entire process described in figure 5.3.

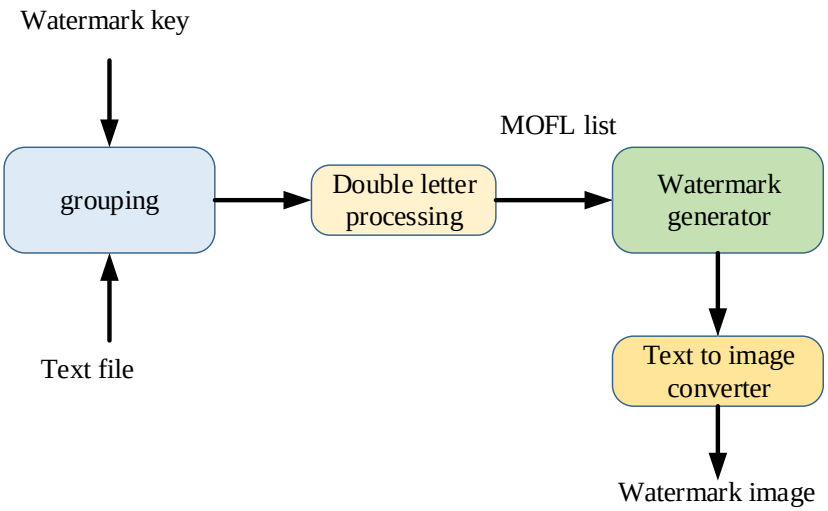

Figure 5.3 General watermark extraction process

5.4 Zero watermarking for text documents

In research field, the text watermarking is emerging technique. The text watermarking algorithms are improved and classified into Image based methods, Syntactic methods, Semantic methods and Structural methods Categories. The binary watermarks are embedded in text image, in image-based methods of text watermarking. The structure of a text is used to embed the watermark in syntactic methods. The watermark embedding is accomplished by using the semantics of text in semantic schemes.

Several algorithms are presented on the basis of these three schemes. Recently, the structural schemes of text watermarking are used which considers the structures of text to embed watermarks. The text is not customized when the watermark is embedded on to it in this scheme. The structural algorithm is not suitable to specific text including poetry, Web contents, legal documents, and the documents which contains mathematical notations with floating point numbers. These types of text watermarking methods are called robust zero watermarking.

The text watermarking techniques are used to secure the contents from illegal copying, forgery, and redistribution. This method also helps to protect texts from the

copyright violations. Moreover, the watermarking technique offers authentication and protection of text documents. The text documents face lots of issues including tampering, copying, reproduction, plagiarism, and paraphrasing attacks. The suitable solution for these issues is digital watermarking which helps in text protection.

The owner of the copyright material can be discovered by using digital watermarking. The original copyright owner of text inputs his/her watermark. The algorithm generates a unique key which correlates to input data. Later, this key is utilized for extraction of watermark, whenever a copyright violation rises in future. Zero watermarking approach is used in which the host text document is not modified to embed watermark, rather the constituents of text are used to generate a key in a unique way to protect it.

The watermarking process has two steps mainly they are; watermark embedding and watermark extraction. The embedding process is done by the original author. Later the extraction process done by certifying authority to confirm ownership and compared the result with the existing algorithms. Embedding Algorithm incorporates the watermark in text. The embedding algorithm logically embeds the watermark in text and produces the author key. The embedded watermark in the text is extracted by extraction algorithm. This process takes the author key got form the certifying authority as input extracts the watermark from the text. The technique kept with the certifying authority that is utilized to solve the copyright violation [85].

5.5 Proposed algorithm for tampering detection

The proposed approach of tampering detection is depends on zero watermarking that uses the features of text contents to create watermark rather than embedding the watermark into the content. This method is based on structural component and word length and it is useable for all types of files. It includes of two steps; text embedding and text extraction. The generation of watermark can be prepared with the owner data and the extraction can be achieved with the CA. The framework of tampering detection technique is shown in figure 5.4.

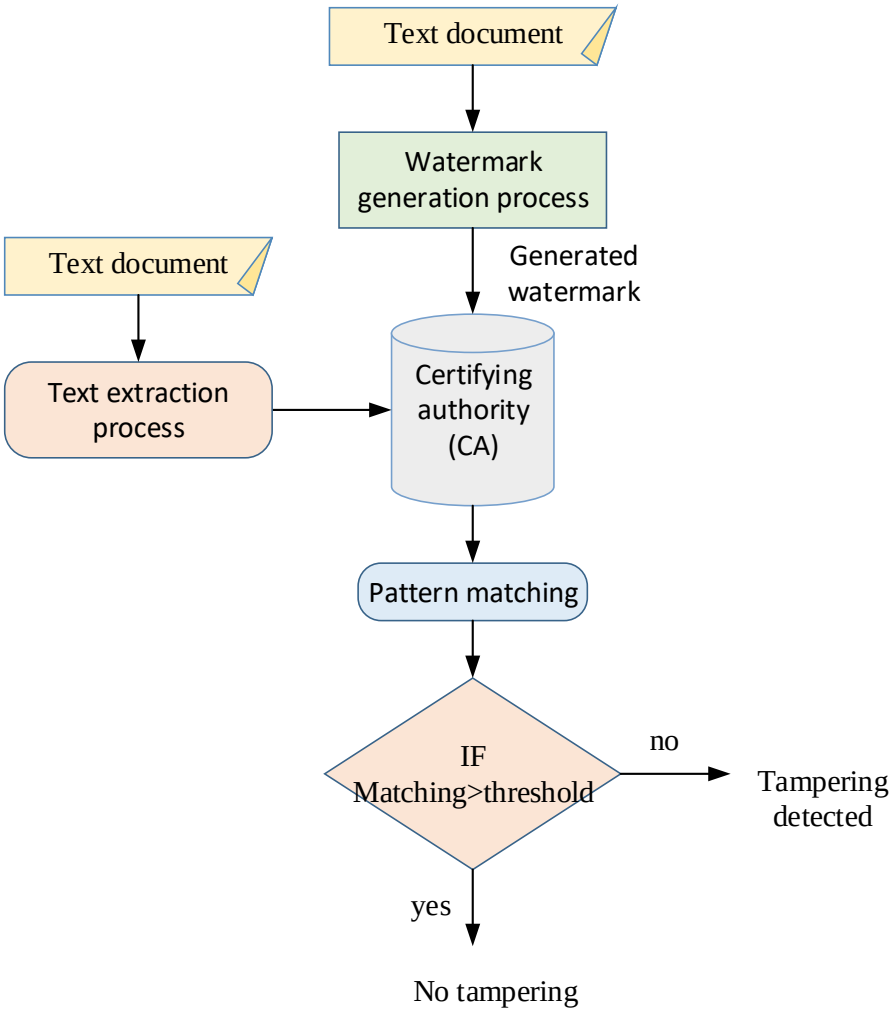

Fig 5.4 overview of proposed HSW based algorithm

In proposed algorithm the watermark is generated for the text document based on HSW method where takes content for tampering detection as an input. Furthermore the extracted pattern of watermark is registered with the certifying authority (CA). The attacker may tamper the content of the file. In the tampering detection process, the extraction algorithm is applied to extract the watermark and the pattern is matched with the registered pattern of CA. The threshold depends on level of pattern matching is set to decide about tampering happened or not.

5.5.1 Watermark generation

The proposed watermarking generation is based on word length and structural components of the text document. Initially, the text partition is formed by considering preposition as a separator. Then the groups are formed by combining the partitions based on the size of the group. In order to form partition and group, the group size and preposition is taken as an input. The frequency of first letter from the double letter word is identified to form the high frequency list. Then the letters are sorted in descending order which places the letter with highest value first. The embedding algorithm is utilized for generating watermark for the text content.

Algorithm 1: HSW based watermark generation
1. Text content T_e and Group size G_s is taken as an input.
2. Partition the text document based on the size of the group G_s .
3. Combine the partitions based on the group size.
4. The number of groups $N_g = \frac{N_P}{G_s}$
5. Find the double letter word and extract the first letter from that word.
6. Compute the frequency d_f of first letter from the double letter word in each group.
7. Based on the frequency of the first letter sort it in descending order.
8. Measure the number of records from the frequency list L_r
9. Find the number of sentences N_s in each group.
10. for a=1 to N_s repeat step 11 to 19
11. Compute the number of words N_w in each sentence.
12. for b=1 to N_w repeat step 13 to 18
13. Compute the number of letter N_l in each word.

14. for c=1 N_l to repeat step 15 to 17

15. if $N_l > 4$

16. L_{t1} = first letter of that word

17 $a = a + 1$

18 $b = b + 1$

19 $c = c + 1$

20. For each partition the first letter is combined to form L_t

21. The total number of L_t is computed for each group.

22. Compare the number of patterns generated with L_r and L_t

23. Eliminate the unwanted patterns for making the number of patterns equal.

24. Joint the pattern which generates the watermark for the text content.

The watermark pattern generated is registered with CA which is considered as a trusted third party of digital community. When clash viewing, the extracted watermark pattern is analyzed with the text content to evaluate its authenticity. The CA executed the detection algorithm and it provides response to the data owner who registered the watermark pattern. The text content may be attacked with several possible ways. The proposed algorithm detects tampering for deletion, insertion and reordering attacks.

5.5.2 Watermark extraction and tampering detection

The extraction algorithm extracts the watermark from the text content. The issue regarding copy right protection can be resolved with CA which keeps the extraction algorithm. Text extraction process is described in algorithm 2.

Algorithm 2: HSW based text extraction and tampering detection
1. Based on the size of the group partition the text extracted. 2. The size of the group is utilized combining the partitions. 3. Consider the double letter word and the word with greater than four letters. 4. Process the consideration for each partition and for each group. 5. Extract the first letter and computes its frequency. 6. The first letter is extracted and combined for each partition if the word length is greater than four. 7. Make the pattern for each partition from step 6. 8. Combine the size of the pattern and extracted first letter from the double letter. 9. Reject the addition information extracted to make unique size. 10. Combine the pattern and the first letter which generates the modified pattern. 11. The extracted pattern is compared with the pattern registered with CA. 12. The score is generated for each pattern based on the pattern similarity. 13. The sore from all patterns is added to get the final score. 14. Set the threshold for the tampering detection. 15. If ($score > threshold$) 16. Document is authenticated. 17. else Tampering detected. 18 Perform the same process for all kinds of attacks.

During extraction, the text document is partitioned based on the preposition which is used in watermark generation. The process of partitioning simplifies the process of

pattern extraction. Since the pattern is extracted for each group separately. The partition is also based on the preposition from the same text contents which utilizes the components of text content. The text group is formed by combining the partitions based on the group size. From each group, the double letter word is analyzed and the frequency for the first letter is computed.

The watermark is extracted and compared with the registered pattern from the CA. For each pattern, the score is generated based on the matching probability. Based on the capacity of matching, the score of all patterns is added to get the combined score. The threshold is set to decide that the tampering is high or low. This algorithm efficiently detects the generated watermark from the text content if there is no attack detected. This kind of text document is termed as authentic text without tampering. The proposed algorithm is detects tampering for the attacks such as insertion, deletion and reordering. Rather than changing the original contents of text document, the characteristics of text is used. The proposed hybrid algorithm has the advantages of high accuracy since it is based on structural components and word length. In addition to that it also authenticates the original text documents.

5.6 Experimental results and analysis

The proposed technique of tampering detection is simulated in MATLAB by varying the number of documents. The absolute score between original watermark pattern and extracted pattern is calculated to evaluate the degree of tampering. The work is evaluated for all possible attacks such as insertion, reordering and deletion. The detection accuracy is considered for the performance evaluation of the proposed method. The performance of the proposed method is compared with the existing approach.

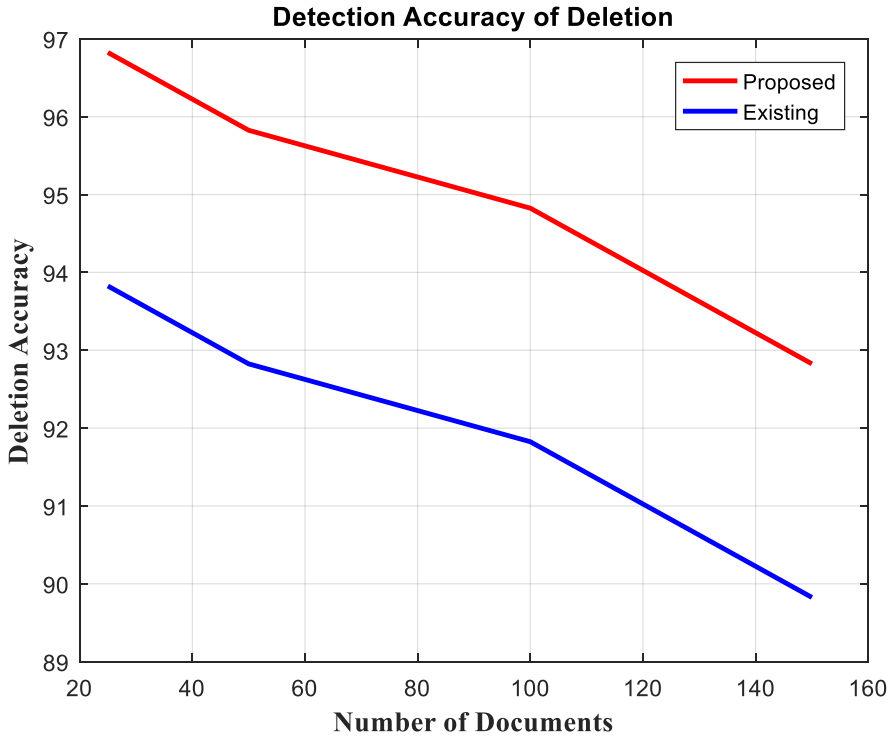

Figure 5.5 Deletion accuracy comparison by varying the number of documents

Figure describes the detection accuracy for deletion attack by changing the number of documents. In this graph, the numbers of documents are varied to calculate the proposed detection performance. The deletion accuracy is evaluated for the number of documents. The detection accuracy for the deletion task is above 97 for the numbers of documents are 20. This detection can be lower than the proposed one for the existing methods. The detection accuracy is decreased by increasing the number of documents. This figure described that the proposed approach has good performance.

Figure describes the detection accuracy for the insertion task. The accuracy of detection is greater than the existing based on the calculation in performance. Based on the changes in number of documents the insertion accuracy will also vary. The insertion accuracy is evaluated for the number of documents. By increasing the number of documents the insertion accuracy is reduced. The proposed method is compared with existing method to discover the performance our work.

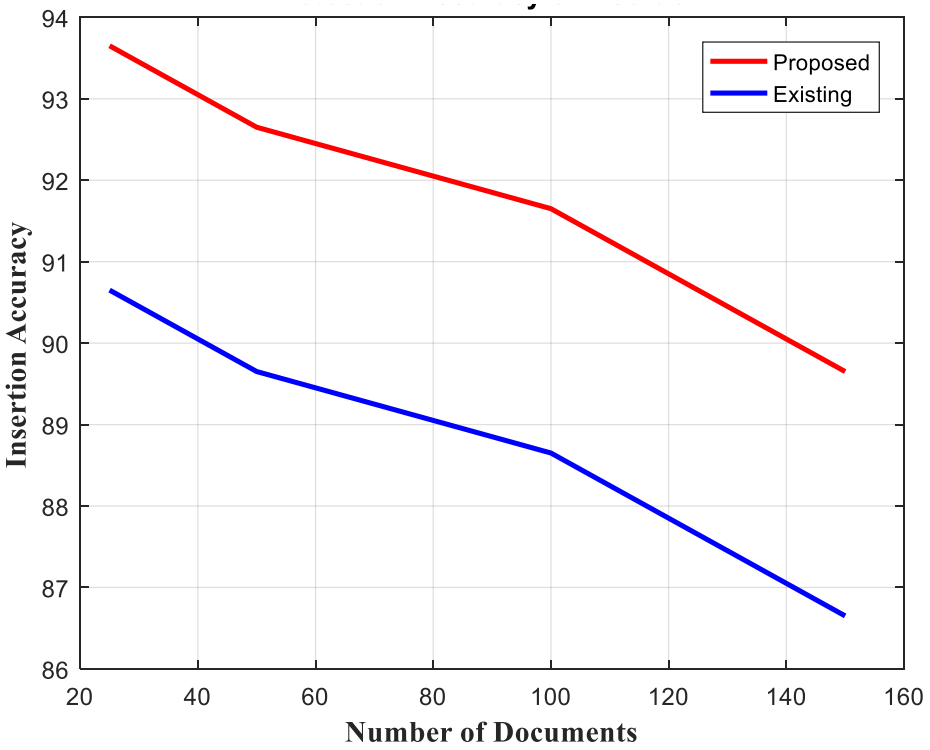

Figure 5.6 Insertion accuracy comparison by varying the number of documents

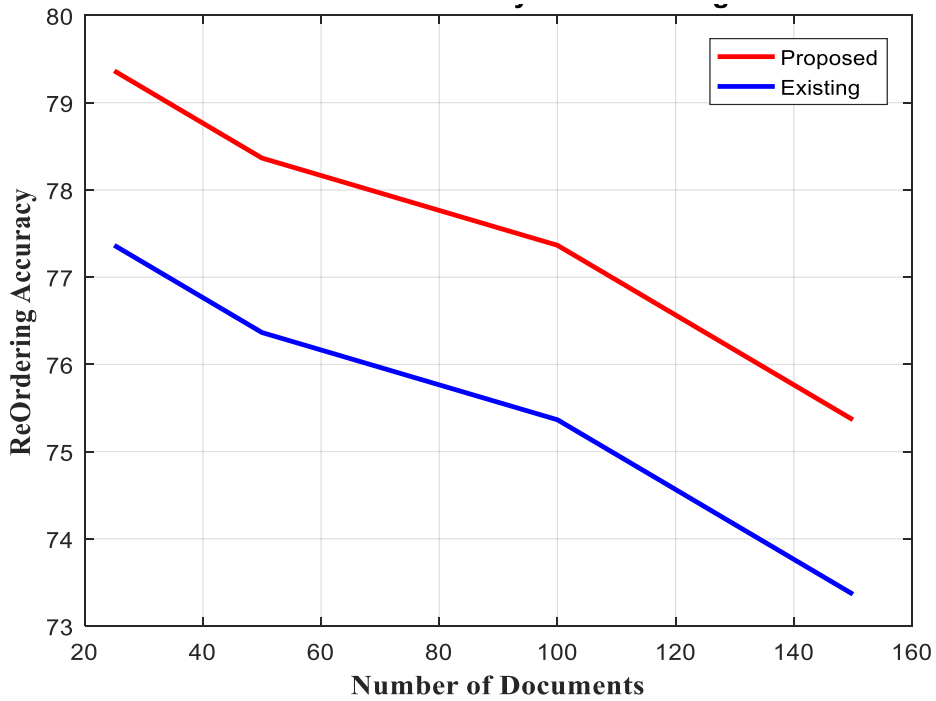

Figure 5.7 Reordering accuracy comparison by varying the number of documents

The detection accuracy for the reordering task described in figure. According to the changes in number of documents the detection accuracy of reordering task also changes. The accuracy is evaluated based on the number of documents. When the numbers of documents are 20, the detection accuracy is 90 and 80. By increasing the number of document the accuracy of the existing approach is reduced.

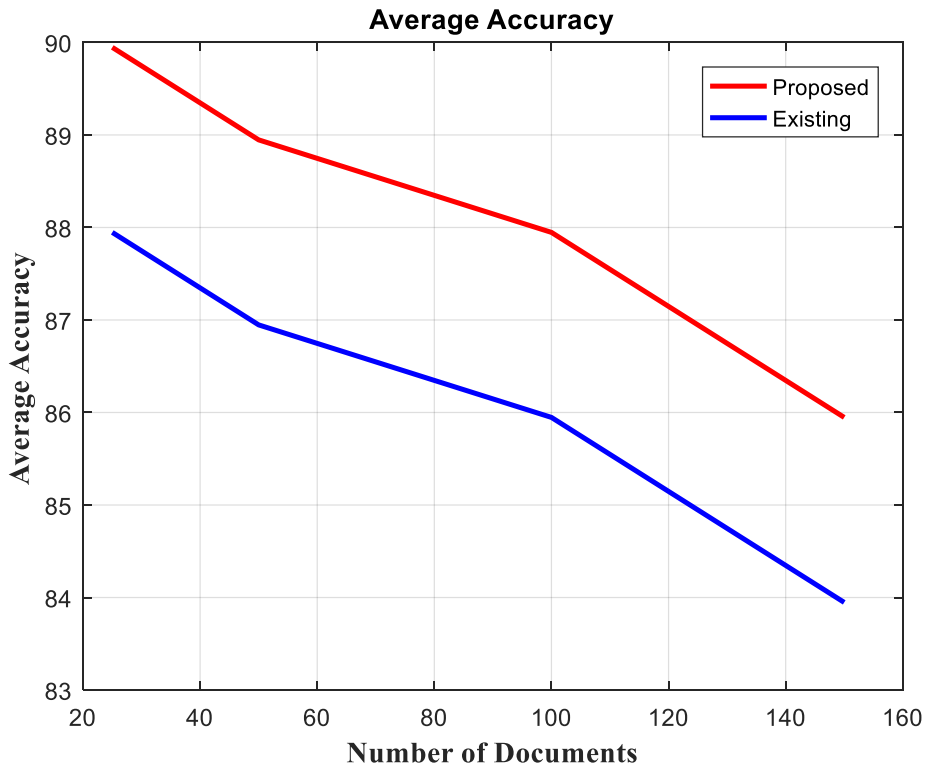

Figure 5.8 Average accuracy comparison by varying the number of documents

The average accuracy for all deletion accuracy for the proposed tampering detection compared with the detection accuracy of existing approaches. The detection performed in this graph by varying the number of documents. The average accuracy for the proposed detection task is shown in above graph. The average accuracy for the existing approach also described.

5.7 Conclusion

Initially, from the data owner the plain text is gained and the content of the text data is analysed. Later, for the plain text watermark is generated with the novel embedding algorithm Hybrid Structural component and word length based zero watermarking. To form groups depends on the group size the text partition is formed and the partition is combined together. Afterwards, the word or letter from each group is analysed to identify the higher occurring list. This list is used for the construction of watermark key on the basis of watermark. The watermark is registered with the certifying authority with the original plain text, author name, date and time. The watermark is generated by the proposed watermarking technique based on the characteristics of plain text rather than embedding the text itself. The text may be attacked or un-attacked after the watermarking. The extraction algorithm for HSW is in which the watermark is extracted from the text and it generates the watermark pattern. After the extraction, for identify the tampering attacks each extracted pattern compared with the pattern registered with the certifying authority.

CHAPTER 6
CONCLUSION AND FUTURE SCOPE

CHAPTER 6

CONCLUSION AND FUTURE SCOPE

This chapter presents the conclusion of this thesis and future work. In first chapter the fundamental theories about e-learning and its evolution, characteristics are provided. The second chapter literature survey provides related researches based on our work. The third chapter provides relevant study about text watermarking and integrity verification. In this chapter overall biometric cryptosystem is explained. Chapter 4 concentrated on the combined Markov model and zero watermarking for integrity verification of PDF English text documents. Chapter 5 presents the proposed HSW based zero watermarking with its simulation results and comparative analysis.

6.1 General summary of the thesis

The increasing of internet usage made a big impact on the society. By the improvement of internet, people send large contents through the web. The digitally stored text documents may goes under the threats including unauthorised copying, modifying, redistribution of the copyrighted text documents, tampering and illegal authentication. To avoid these security threats, watermarking techniques are utilized. The watermarking techniques prevent from the illegal attacks. In our thesis HSW based watermarking is proposed. In each chapter the concept of the watermarking is briefly explained. The short summary of the thesis is described as follows.

In chapter 1, the introduction for the e-learning is discussed with its requirements and applications. Then the development and growth of the e-learning is briefly explained. The evolution of the e-learning method is described by block diagram. The factors of e-learning development are discussed and explained. The considerations for developing effective e-learning and three-tier technology use model discussed under e-learning development and growth. The e-learning overview and structure is described with block diagram. The concept of E-learning system components also described with diagram. The user expectations in e-learning are briefly discussed in this chapter. The challenges and security threats occurring in e-learning is explained.

The need of e-learning security and applications of the e-learning are discussed. Integrity verification techniques and integrity attacks are described. The advantages of e-learning is analysed in this chapter.

In chapter 2, the review for the existing approaches is presented with its disadvantages and the requirement of proposed approaches. The performances of e-learning methods from various literatures are discussed and the comparisons of these methods are provided. Then e-learning system challenges are provided. The existing watermarking techniques are discussed with its disadvantages and the algorithms used. Then the zero watermarking methods are analysed with its disadvantages and the different algorithms used. The text content based zero watermarking methods also discussed in this chapter with different existing approaches.

In chapter 3, text watermarking and integrity verification is mainly concentrated. Initially the introduction for the text watermarking and integrity verification provided. Then the classification of the watermarking methods briefly explained with diagrams. In this chapter all categorization of the watermarking methods are described. Then the stages of watermarking are discussed with block diagram. The explained watermarking stages are embedding process, extraction process and detection process. Then the possible attacks may occur in these stages is explained. After that error correcting output codes is analysed with diagram and algorithms. The biometric cryptosystem structure is discussed with diagram. The biometric system has two phases, enrolment phase and authentication phase. There algorithms are also discussed in this chapter. This chapter briefly describes about biometric cryptosystem.

In chapter 4, the combined Markov model and zero watermarking for integrity verification of PDF English text documents is described. In this chapter zero watermarking based techniques for integrity verification are analysed and discussed. The embedding process and extraction process of the zero watermarking methods are discussed with block diagrams. Then zero watermarking algorithm based on multiple occurrences of letters also discussed with embedding and extraction algorithms. After that the brief description of combined Markov model based watermarking is provided.

Then watermark generation and detection process analysed with diagram and algorithms. In this chapter LNMZW3 and WNMZW5 algorithm is presented and the compared its performance with other existing algorithms. The watermark embedding and extraction is carried out by this LNMZW3 and WNMZW5 algorithm. The simulation results provided that the LNMZW approach has higher performance and robustness against common volumes of insertion, deletion and reorder attacks. The presented LNMZW3 and WNMZW5 algorithm is more secure than the other existing algorithms. The watermark robustness is improved with short and medium document size.

In chapter 5, our proposed HSW based zero watermarking is presented and analysed the performance. Zero watermarking based on structural components is discussed. The detailed description of watermarking process for text document is provided. In this chapter HSW based algorithm is proposed for tampering detection. The advantages present in HSW based algorithm also discussed. The simulation results of the proposed method is compared with existing method and analysed the performance of proposed HSW based tampering detection. The experiment is done in different attacks like insertion, deletion and reorder attacks. According to the number of documents the accuracy of attacks also changed. The proposed method performance achieves good performance when compared to existing works.

6.2 Results and conclusions

In this thesis, the efficient integrity verification and tampering detection method is presented with HSW based watermarking algorithm. Various e-learning systems and their basics are studied in this thesis. The combined Markov model and zero watermarking technique studied. Then the simulation results of LNMZW3 and WNMZW5 are compared with the algorithms WNMZW4, LNMZW1 and LNMZW2 under insertion, deletion and reorder attacks. The text watermarking methods and integrity verification of the documents with various methods are analysed. The proposed HSW based watermarking technique is analysed and compared with existing method. The simulation results of the proposed method is compared with existing

method and analysed the performance of proposed HSW based tampering detection. The proposed method results are compared and analysed under insertion, deletion and reorder attacks.

6.3 Future work

In this work, the accuracy for insertion, deletion and reorder attacks of text document are investigated by proposed HSW based watermarking technique. The performance of the proposed method is better when compared to other works. More efficient techniques are required for good accuracy. Moreover we can use latest NLP techniques instead of Markov model for text analysis and generate pattern key. The proposed (LNMZW) and (WNMZW) approaches can we develop it as a combined with owner Meta data and apply them not only for text content but can extend to another media like image, audio and video, and we have plan to improve our algorithms in proposed approaches to better. Thus, the future work of this thesis concentrates on improving more techniques for tampering detection and integrity verification in English and Arabic text documents. Furthermore I hope to improve my approach for other e-services such as e-commerce, e-government, and e-banking.

REFRENCES

REFRENCES

[1] Hassanzadeh, Alireza, FatemehKanaani, and ShábanElahi. "A model for measuring e-learning systems success in universities." Expert Systems with Applications 39.12 (2012): 10959-10966.

[2] Kaur, Eep, and KewalKrishan. "Cluster Analysis of Behavior of E-learners." (2013).

[3] Aparicio, Manuela, Fernando Bacao, and Tiago Oliveira. "Cultural impacts on e-learning systems' success." *The Internet and Higher Education* 31 (2016): 58-70.

[4] Lee, Yi-Hsuan, Yi-Chuan Hsieh, and Chun-Yuan Ma. "A model of organizational employees'e-learning systems acceptance." *Knowledge-based systems* 24.3 (2011): 355-366.

[5] Conole, Grainne, Janice Smith, and Su White. "A critique of the impact of policy and funding." *Contemporary perspectives in e-learning research: Themes, methods and impacts on practice* (2007): 38-54.

[6] Rabuzin, Kornelije, Miroslav Baca, and Mario Sajko. "E-learning: Biometrics as a Security Factor." *Computing in the Global Information Technology, 2006. ICCGI'06. International Multi-Conference on*. IEEE, 2006.

[7] Alwi, Najwa Hayaati Mohd, and Ip-Shing Fan. "E-learning and information security management." *International Journal of Digital Society (IJDS)* 1.2 (2010): 148-156.

[8] Liaw, Shu-Sheng. "Investigating students' perceived satisfaction, behavioral intention, and effectiveness of e-learning: A case study of the Blackboard system." *Computers & education* 51.2 (2008): 864-873.

[9] Al-Ajlan, Ajlan S. "E-Learning Certificate Using Digital Watermarking Technology." *IOSR Journal of Computer Engineering*. 16 (4). pp. 81–93(2014).

[10] Paechter, Manuela, Brigitte Maier, and Daniel Macher. "Students' expectations of, and experiences in e-learning: Their relation to learning achievements and course satisfaction." *Computers & education* 54.1 (2010): 222-229.

[11] Ghobadi, A. R., et al. "How watermarking secures e-learning system." *e-Learning and e-Technologies in Education (ICEEE), 2012 International Conference on*. IEEE, 2012.

[12] Barik, Nikhilesh, and Sunil Karforma. "Risks and remedies in e-learning system." *arXiv preprint arXiv:1205.2711* (2012).

[13] Cardenas, Roberto Gomez, and Erika Mata Sanchez. "Security challenges of distributed e-learning systems." *International Symposium and School on Advancex Distributed Systems*. Springer, Berlin, Heidelberg, 2005.

[14] Ismail, Johan. "The design of an e-learning system: Beyond the hype." *The internet and higher education* 4.3-4 (2001): 329-336.

[15] Kim, Won-gyum, and HeungKyu Lee. "Multimodal biometric image watermarking using two-stage integrity verification." *Signal Processing* 89.12 (2009): 2385-2399.

[16] Sajjadi, Zahra, Ali asghar Khodami, and Nasser Modiri. "Learning Contents integrity verification on E-Learning Systems Using Digital Watermarking Technique." *Information and Communication Technologies: From Theory to Applications, 2008. ICTTA 2008. 3rd International Conference on*. IEEE, 2008.

[17] Luchev, Detelin. "Bulgarian National Ethnographic Museum–Meeting the Challenges of Digitisation." (2007).

[18] Zhang, Dongsong, and Jay F. Nunamaker. "Powering e-learning in the new millennium: an overview of e-learning and enabling technology." *Information systems frontiers* 5.2 (2003): 207-218.

[19] Truong, Huong May. "Integrating learning styles and adaptive e-learning system: Current developments, problems and opportunities." Computers in human behavior 55 (2016): 1185-1193.

[20] Al-Gahtani, Said S. "Empirical investigation of e-learning acceptance and assimilation: A structural equation model." Applied Computing and Informatics 12.1 (2016): 27-50.

[21] Harrati, Nouzha, et al. "Exploring user satisfaction for e-learning systems via usage-based metrics and system usability scale analysis." Computers in Human Behavior 61 (2016): 463-471.

[22] Dong, Bo, et al. "An e-learning ecosystem based on cloud computing infrastructure." Advanced Learning Technologies, 2009. ICALT 2009. Ninth IEEE International Conference on. IEEE, 2009.

[23] Aparicio, Manuela, Fernando Bacao, and Tiago Oliveira. "Cultural impacts on e-learning systems' success." The Internet and Higher Education 31 (2016): 58-70.

[24] Christudas, Beulah Christalin Latha, E. Kirubakaran, and P. Ranjit Jeba Thangaiah. "An evolutionary approach for personalization of content delivery in e-learning systems based on learner behavior forcing compatibility of learning materials." *Telematics and Informatics* (2017).

[25] Shee, Daniel Y., and Yi-Shun Wang. "Multi-criteria evaluation of the web-based e-learning system: A methodology based on learner satisfaction and its applications." *Computers & Education* 50.3 (2008): 894-905.

[26] Tao, Yu-Hui. "Typology of college student perception on institutional e-learning issues–An extension study of a teacher's typology in Taiwan." *Computers & Education* 50.4 (2008): 1495-1508.

[27] Ong, Chorng-Shyong, Jung-Yu Lai, and Yi-Shun Wang. "Factors affecting engineers' acceptance of asynchronous e-learning systems in high-tech companies." *Information & management* 41.6 (2004): 795-804.

[28] Govindasamy, Thavamalar. "Successful implementation of e-learning: Pedagogical considerations." *The internet and higher education* 4.3-4 (2001): 287-299.

[29] Wang, Yi-Shun, Hsiu-Yuan Wang, and Daniel Y. Shee. "Measuring e-learning systems success in an organizational context: Scale development and validation." *Computers in Human Behavior* 23.4 (2007): 1792-1808.

[30] Luo, Nuan, Mingli Zhang, and Dan Qi. "Effects of different interactions on students' sense of community in e-learning environment." *Computers & Education* 115 (2017): 153-160.

[31] Mohammadyari, Soheila, and Harminder Singh. "Understanding the effect of e-learning on individual performance: The role of digital literacy." *Computers & Education* 82 (2015): 11-25.

[32] Coatrieux, Gouenou, et al. "A watermarking-based medical image integrity control system and an image moment signature for tampering characterization." *IEEE journal of biomedical and health informatics* 17.6 (2013): 1057-1067.

[33] Singh, Priyanka, and Suneeta Agarwal. "A self recoverable dual watermarking scheme for copyright protection and integrity verification." *Multimedia Tools and Applications* 76.5 (2017): 6389-6428.

[34] Goyal, Leena, et al. "A robust method for integrity protection of digital data in text document watermarking." *Int. J. Sci. Res. Dev* 1.6 (2014): 14-18.

[35] Jalil, Zunera, and Anwar M. Mirza. "Text watermarking using combined image-plus-text watermark." *Education Technology and Computer Science (ETCS), 2010 Second International Workshop on*. Vol. 1. IEEE, 2010.

[36] Zhou, Yaxun, and Wei Jin. "A novel image zero-watermarking scheme based on DWT-SVD." Multimedia Technology (ICMT), 2011 International Conference on. IEEE, 2011.

[37] Jalil, Zunera, et al. "A zero text watermarking algorithm based on non-vowel ASCII characters." *Educational and Information Technology (ICEIT), 2010 International Conference on*. Vol. 2. IEEE, 2010.

[38] Yingjie, Meng, et al. "A Zero-Watermarking Scheme for Prose Writings." *Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), 2017 International Conference on*. IEEE, 2017.

[39] Feng, Gui, and Xihui Huang. "An improved DCT based zero-watermarking algorithm for text image." *Anti-Counterfeiting, Security and Identification (ASID), 2012 International Conference on*. IEEE, 2012.

[40] Singh, Anushikha, et al. "An SVD based zero watermarking scheme for authentication of medical images for tele-medicine applications." *Telecommunications and Signal Processing (TSP), 2016 39th International Conference on*. IEEE, 2016.

[41] Rani, Asha, et al. "A zero-watermarking scheme using discrete wavelet transform." *Procedia Computer Science* 70 (2015): 603-609.

[42] Tsai, Hung-Hsu, Yen-Shou Lai, and Shih-Che Lo. "A zero-watermark scheme with geometrical invariants using SVM and PSO against geometrical attacks for image protection." *Journal of Systems and Software* 86.2 (2013): 335-348.

[43] Jalil, Zunera, Anwar M. Mirza, and Maria Sabir. "Content based zero-watermarking algorithm for authentication of text documents." *arXiv preprint arXiv:1003.1796* (2010).

[44] Ba-Alwi, Fadl M., Mokhtar M. Ghilan, and Fahd N. Al-Wesabi. "Content authentication of English text via internet using zero watermarking technique and Markov model." *International Journal of Applied Information Systems (IJAIS)* 7.1 (2014): 25-36.

[45] Meng, Yingjie, et al. "Chinese text zero-watermark based on sentence's entropy." *Multimedia Technology (ICMT), 2010 International Conference on*. IEEE, 2010.

[46] Al-Wesabi, Fahd N., Adnan Z. Alshakaf, and Kulkarni U. Vasantrao. "A zero text watermarking algorithm based on the probabilistic weights for content authentication of text documents." *Proceedings on National Conference on Recent Trends in Computing NCRTC*. 2012.

[47] Feng, Gui, and Xihui Huang. "An improved DCT based zero-watermarking algorithm for text image." *Anti-Counterfeiting, Security and Identification (ASID), 2012 International Conference on*. IEEE, 2012.

[48] Jaseena, K. U., and Anita John. "An invisible zero watermarking algorithm using combined image and text for protecting text documents." *International Journal on Computer Science and Engineering* 3.6 (2011): 2265-2272.

[49] Tayan, Omar, Y. Alginahi, and Muhammad N. Kabir. "Performance assessment of zero-watermarking techniques for online Arabic textual-content." *Life Science Journal* 10.4 (2013): 93-100.

[50] Qi, Xitong, and Yuling Liu. "Cloud model based zero-watermarking algorithm for authentication of text document." *Computational Intelligence and Security (CIS), 2013 9th International Conference on*. IEEE, 2013.

[51] Kim, Young-Won, Kyung-Ae Moon, and Il-Seok Oh. "A Text Watermarking Algorithm based on Word Classification and Inter-word Space Statistics." *ICDAR*. 2003.

[52] Jaseena, K. U., and Anita John. "Text watermarking using combined image and text for authentication and protection." *International Journal of Computer Applications* 20.4 (2011): 8-13.

[53] Zhu, Ping, et al. "A text zero-watermarking algorithm based on Chinese phonetic alphabets." *Wuhan University Journal of Natural Sciences* 21.4 (2016): 277-282.

[54] Singh, Priyanka, and Suneeta Agarwal. "A self recoverable dual watermarking scheme for copyright protection and integrity verification." *Multimedia Tools and Applications* 76.5 (2017): 6389-6428.

[55] Mali, Makarand L., Nitin N. Patil, and J. B. Patil. "Implementation of text watermarking technique using natural language watermarks." *Communication Systems and Network Technologies (CSNT), 2013 International Conference on*. IEEE, 2013.

[56] Singh, Amit Kumar, Mayank Dave, and Anand Mohan. "Hybrid technique for robust and imperceptible multiple watermarking using medical images." *Multimedia Tools and Applications*75.14 (2016): 8381-8401.

[57] KO, Mohammed Aarif. "A Study of Digital Watermarking Techniques."

[58] Pik-Wah, C. H. A. N. "Digital video watermarking techniques for secure multimedia creation and delivery." *The Chinese University of Hong Kong* (2004).

[59] Chawla, Gaurav, Ravi Saini, and Rajkumar Yadav. "Classification of watermarking based upon various parameters." *International Journal of Computer Applications & Information Technology* 1.II (2012).

[60] Shih, Frank Y. *Digital watermarking and steganography: fundamentals and techniques*. CRC press, 2017.

[61] Guru, Jaishri, and Hemant Damecha. "Digital watermarking classification: survey." *International Journal of Computer Science Trends and Technology (IJCST) vol* 5 (2014): 8-13.

[62] Bharati, Patil Devidas, and Patil Namdeo Nitin. "Text watermarking algorithm using structural approach." *Information and Communication Technologies (WICT), 2012 World Congress on*. IEEE, 2012.

[63] Premaratne, P., and C. C. Ko. "A novel watermark embedding and detection scheme for images in DFT domain." (1999): 780-783.

[64] Radharani, S., and M. L. Valarmathi. "A study on watermarking schemes for image authentication." *International Journal of Computer Applications* 2.4 (2010): 24-32.

[65] Nematollahi, Mohammad Ali, Chalee Vorakulpipat, and Hamurabi Gamboa Rosales. "Text Watermarking." *Digital Watermarking*. Springer, Singapore, 2017. 121-129.

[66] Kaur, Manmeet, and Kamna Mahajan. "An existential review on text watermarking techniques." *International Journal of Computer Applications* 120.18 (2015).

[67] Kamaruddin, Nurul Shamimi, et al. "A Review of Text Watermarking: Theory, Methods, and Applications." *IEEE Access* 6 (2018): 8011-8028.

[68] Rathgeb, Christian, and Andreas Uhl. "A survey on biometric cryptosystems and cancelable biometrics." *EURASIP Journal on Information Security* 2011.1 (2011): 3.

[69] Nazari, Sara, Mohammad-Shahram Moin, and Hamidreza Rashidy Kanan.

Computers & Electrical Engineering (2018).

[70] Kaur, Harkeerat, and Pritee Khanna. "Biometric template protection using cancelable biometrics and visual cryptography techniques." *Multimedia Tools and Applications*75.23 (2016): 16333-16361.

[71] Wang, Wei, et al. "A novel robust zero watermarking scheme based on DWT and SVD." *Image and Signal Processing (CISP), 2011 4th International Congress on*. Vol. 2. IEEE, 2011.

[72] Khan, Asifullah, Anwar M. Mirza, and Abdul Majid. "Optimizing perceptual shaping of a digital watermark using genetic programming." *Iranian journal of electrical and computer engineering* 3.2 (2004): 144.

[73] Tayan, Omar, Muhammad N. Kabir, and Yasser M. Alginahi. "A hybrid digital-signature and zero-watermarking approach for authentication and protection of sensitive electronic documents." *The Scientific World Journal* 2014 (2014).

[74] Cao, Hanqiang, et al. "A zero-watermarking algorithm based on DWT and chaotic modulation." *Independent Component Analyses, Wavelets, Unsupervised Smart Sensors, and Neural Networks IV*. Vol. 6247. International Society for Optics and Photonics, 2006.

[75] Yang, Yu, et al. "A Novel Robust Zero-Watermarking Scheme Based on Discrete Wavelet Transform." *Journal of Multimedia* 7.4 (2012).

[76] Jalil, Zunera, et al. "Improved zero text watermarking algorithm against meaning preserving attacks." *World academy of science, engineering and technology* 46 (2010): 592-596.

[77] Kaur, Sukhpreet, and Geetanjali Babbar. "A Zero-Watermarking algorithm on multiple occurrences of letters for text tampering detection." *International Journal on Computer Science and Engineering* 5.5 (2013): 294.

[78] Ba-Alwi, Fadl M., Mokhtar M. Ghilan, and Fahd N. Al-Wesabi. "Content authentication of English text via internet using zero watermarking technique and Markov model." *International Journal of Applied Information Systems (IJAIS)* 7.1 (2014): 25-36.

[79] Wang, K., Babenko, B. and Belongie, S., 2011, November. End-to-end scene text recognition. *In Computer Vision (ICCV), IEEE International Conference on* (2011) pp. 1457-1464.

[80] Maria Chroni, and Stavros D. Nikolopoulos. "Watermarking PDF Documents using Various Representations of Self-inverting Permutations." *arXiv preprint arXiv:1501.02686*(2015).

[81] Bindra, Gundeep Singh. "Invisible Communication through Portable Document File (PDF) Format." *Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP), 2011 Seventh International Conference on*. IEEE, 2011.

[82] Leena Goyal, Manoj Raman, PrateekDiwan, Mukesh Vijay, Leena Goyal, Manoj Raman, PrateekDiwan, and Mukesh Vijay. "A robust method for integrity protection

of digital data in text document watermarking." *Int. J. Sci. Res. Dev* 1, no. 6 (2014): 14-18.

[83] Reem A. Alotaibi, and Lamiaa A. Elrefaei. "Improved capacity Arabic text watermarking methods based on open word space." *Journal of King Saud University-Computer and Information Sciences* (2017).

[84] Bharati, Patil Devidas, and Patil Namdeo Nitin. "Text watermarking algorithm using structural approach." *Information and Communication Technologies (WICT), 2012 World Congress on*. IEEE, 2012.

[85] Bhambri, Pankaj, and Pradeep Kaur. "A Novel Approach of Zero Watermarking for Text Documents." *International Journal of Ethics in Engineering & Management Education (IJEEE)* 1.1 (2014): 34-38.

[86] N. S. Kamaruddin, A. Kamsin, L. Y. Por, and H. Rahman, "A Review of Text Watermarking : Theory , Methods and Applications," *IEEE Access*, vol. 3536, no. c, 2018.

AUTHOR'S BIOGRAPHY

Mr. Fatek Saeed holds Bachelor of Computer Science from Thamar University and Master of Computer Information Systems from Arab Academy Jordon, Sanaa branch. His interests include E-Learning, E-Commerce, E-Business, E-Banking, E-Government, Artificial Intelligence, Data Mining, Information System Management, Information Security, and Systems Analysis. He has increased his interests in Strategies development, Accreditation & Quality Assurance, Higher Education Management, Institutions Evaluation and Monitoring during his job in Ministry of Higher Education and Scientific Research at Yemen where last his job there before get scholarship for PhD was an office manager of Deputy Minster. He has rich knowledge of Accreditation & Quality Assurance where he participates in many important international conferences and workshops. Furthermore, he has a lot of international certificates regarding that and for Human Resource Management and Interpersonal Behavior from Nuffic, Maastricht School of Management, University of Twente, and Delft University of technology Holland universities and so on. He has officially travelled to Syria, Jordon, Philippines, UAE, KSA, India and Oman.