بسم الله الرحمن الرحيم

Islamic University – Gaza

Deanery of Post Graduate Studies

Faculty of Information Technology

الجامعة الاسلامية – غزة

عمادة الدراسات العليا

كلية تكنولوجيا المعلومات

New Method to Detect Text Fabrication in Scanned Documents

Prepared By

Fadi H. Naser Hasan

120100576

Supervisor*:*

Dr. Tawfiq Barhoom

Submitted in partial fulfillment of the requirements of the degree of Master In Information Technology

1436/2015

Abstract

In our life, we rely on the documents and papers almost in all transaction areas, and with the importance of these documents, there is a new crime appeared in the society called, documents forgery, and it is one of the most serious crimes in societies.

Many researchers try to find ways to detect that forgery in order to prevent these danger crimes so they use some methods and approaches to enhance the performance of detection, but with variance performance between their results depend on the difference between their methods.

This research introduces a new method, which has two stages and five processes to detect the forgery on scanned document. The first three processes using to preprocessing the documents to them suitable for the next two stages, the fourth process is included in stage one, which using to Extract Max Frequency Intensity feature (EMFI) from the documents pixels, and the fifth process is included in stage two, which using to Extract Edge Gradient (EEG) to find the variance between the real text and the fabricated text. The final process is the serial combination, coloring and locating the suspected pixels.

Moreover, we built our own dataset by fabricating some official documents from Islamic University in Gaza (IUG) to test and evaluate the performance of our method, and we measured the system performance for the serial combining between EMFI and EEG methods, by using the recall and precision from the confusion matrix

Keywords: Document Forgery, Fabricated Documents, Information Security.

عنوان البحث :

طريقة جديدة للكشف عن تزوير النصوص في المستندات الممسوحة ضوئيا

الملخص:-

في حياتنا اليومية، نعتمد على المستندات والأوراق تقريبا في جميع المعاملات، ومع أهمية استخدامات هذه الوثائق، ظهرت هناك جريمة جديدة في المجتمع، ألا وهي تزوير هذه الوثائق، وبالتالي أصبحت هذه الجريمة واحدة من أخطر الجرائم في المجتمع.

ولأجل منع مثل هذه الجرائم ودرء خطرها حاول العديد من الباحثين إيجاد طرق للكشف عن التزوير باستخدام بعض الطرق والأساليب والاجتهادات التي استخدموها لتعزيز الدقة في الكشف عن هذا التزوير في المستندات، ولكن تبين وجود تباين بين نتائجهم التي ظهرت, ويفسر وجود هذا التباين بناء على الفرق بين الأساليب المستخدمة والمقترحة من قبلهم.

هذه الرسالة استخدمت طريقة جديدة للكشف عن التزوير في المستندات الممسوحة ضوئيا تتكون مرحلتين خمس عمليات رئيسية,العمليات الثلاثة الأولى تستخدم لتهيئة المستند ليكون ملائما للعمل عليه في المرحلتين التاليتين, العملية الرابعة مدرجة ضمن المرحلة الأولى وتستخدم لاستخراج أكثر تكرار كثافة من البكسلات المكونة للمستند, أما العملية الخامسة فهي مدرجة ضمن المرحلة الثانية وتستخدم لاستخراج انحدار تدرج اللون عند حافة الكلمات للعثور على التباين بين النص الملفق والنص الحقيقي, العملية الأخيرة هي عملية الدمج والتلوين وتحديد الموقع للبكسلات المشكوك بها.

وعلاوة على ذلك، قمنا ببناء مجموعة البيانات الخاصة بنا من خلال افتعال بعض الوثائق الرسمية من الجامعة الإسلامية في غزة (الجامعة الإسلامية) لاختبار وتقييم أداء أسلوبنا، وقمنا بقياس أداء النظام بعد الدمج بين طريقة استخراج كثافة اللون الأسود وطريقة استخراج الانحدار في تدرج اللون عند حواف الكلمات، وذلك باستخدام الاستدعاء ودقة التوقع من مصفوفة الارتباك.

الكلمات الدالة: تزوير الوثائق , المستند المزور , أمن المعلومات

الإهداء

الى الذين قال فيهم الله تعالى (واخفض لهما جناح الذل من الرحمة وقل رب ارحمهما كما ربياني صغيرا)

الى القدوة في داخلي . . . نبع العطاء الذي لم يحرمني شيئا . . . من زرع الأخلاق والطيبة في داخلي . . . الى أبي الطيب أطال الله عمره

الى الزهرة التي لا تذبل . . . الى من تحملت لأجلنا الكثير . . . الى من يسكن الكون عند سماع اسمها . . . الى أمي

الى اخوتي واخواتي . . . سندي في هذه الدنيا

الى الذين تحملوني واحتضنوني وتحملوا تغير أمزجتي وضغوطي . . . الى زوجتي وأبنائي

الى رفاق الدرب . . . الى من ساعدني في بناء المستقبل . . . الى أروع وأنبل البشر . . . الى من وقفوا معي في مسيرتي

الى من رفعوا رايات العلم . . . الى من علمني حرفا . . . أساتذتي الكرام

وأخص بالذكر الأستاذ الدكتور / توفيق برهوم . . . الرائع دائما

أهدي هذا الجهد المتواضع الى كل من قال لا اله الا الله محمد رسول الله

أهدي لكل من يقرأ رسالتي تعبي وجهدي وسهري

Table of Contents

List of Figures

List of Tables

List of Abbreviations

(EMFI)	Extracting Max Frequency Intensity
(EEG)	Extracting Edge Gradient
(FDS)	Forgery Detection System
(ROI)	Region Of Interest
(QOU)	Quds Open University
(IUG)	Islamic University Gaza

Chapter 1

Introduction

1.1 Introduction

While digital photocopy documents are widely used in our life, especially in official and business environment, many types of forgery began to appear with different techniques, these techniques are classified into three general categories, copy/move forgery, image splicing and image retouching [35]. These types of digital forgeries are mainly performed on data, documents, cheques, images, and video. However, there are some differences between them.

The copy/move forgery is one of the most famous type of image forgery whereas a part of the image is copied and pasted on another part at the same image.

The image splicing is almost the same as the copy/move but here the fabricator copy or cutting a part of another image and pastes it in the target image.

The retouching means to make some retouch in the image by changing or adding or deleting some contents of the image.

Each kind of manipulations that changes the content of any documents is illegal [29].

Although, not all these techniques are commonly used in document forgery, the fabrication of the documents is one of retouching technique that is used in documents forgery [35]. The retouching is mostly used to enhance or reduce the document or the image features. Usually this type of forgery is being realized by changing the color, texture or intensity of the objects; or simply introducing some blur for defusing the objects [35], all that fabrications techniques will produce a false or fabricated document.

A False or Fabricated document is a technique employed to create verisimilitude in a work of fiction. By inventing and inserting documents that appear to be factual, an author tries to create a sense of authenticity beyond the normal and expected suspension of disbelief for a work of art. The fabricated photocopy documents are generated to gain some short term or long term benefits unlawfully. This poses a serious threat to the system and the economics of a nation. In general, such frauds are noticed in the application areas where photocopy documents are just enough [31].

For examples, the false-document may include at the following: Fake police reports, newspaper articles, bibliographical references, documentary footage, or using the legal names of performers or writers in a fictional context. Supplementary material such as badges, identity cards (IC), diaries, letters or artifacts can also be included, and this extends the exercise beyond the confines of the text. So the **F**abricated documents intentionally blur the boundaries between fiction and fact.

Several researches try to find a solution for the document fabrication problem by building and proposing some models, approaches and techniques to detect the fabrication from documents by using a verity of features that extracted from image document [7,9,14,29,30] such as the line alignment information, and the spaces between characters in English language letters and also with some Pixels properties.

However, there are a lot of challenges in this domain, one of them is still the major challenge which can meet document forensics, and this challenge is to find completely the actual location of fabrication on the document [29]. Until this time this problem remains unsolved completely [29], therefor we try to mitigate it by our new method.

Also, there is no public data set available for experimentation [15] until now. Hence, for the purpose of experimentation a considerable size of data samples for testing must be collected. These samples may include conference certificates, official

documents, birth certificates, death certificates, degree certificates and transfer certificates, therefor, in our dataset we use an official documents from Islamic University in Gaza (IUG) as main dataset and also we evaluate our system be using one school certificate from one of the UNRWA schools, and one University degree certificate from Al Quds Open University (QOU).

Finally, it is well known that the best solution to detect fabrication in any image or document can be achieved by comparing the fabricated image with the original one, if we can get it.

However the problem becomes more complex if we cannot bring the original image, so here we have a blind document [35] , and this is the major challenge to find if this document is fabricated or not?.

In this research we implemented a new method to address the image documents fabrication problem, depending on extracting features from the image, such as: the max frequency intensity of pixels and edge gradient, to find some variance between the fabricated text and the original to detect and locate the fabrication.

1.2 Problem Statement:

The research problem is the fabrication in text in the official scanned documents, and the weaknesses performance in the systems that detect documents fabrication.

1.3 Objectives:

The main and the specific objectives of the thesis :

1.3.1 Main objective:

Propose a new method based on determining the features of (intensity frequency and edge gradient) from the document's pixels to detect fabrication in scanned documents. Also, to enhance the performance of systems that depend on the pixels properties to detect documents fabrication.

1.3.2 Specific objectives:

- Explore different types of forgeries and fabrication in documents to learn more about the characteristics of these types, and how to detect.
- Explain some useful pre-processing on images, which can help to extract the features such as threshold, noise removal.
- Design and implement the new method based on determining the features (intensity and edge gradient) from the document's pixels to detect fabrication in scanned documents
- Collect a set of forged documents for use in testing and evaluating the method from friends such as child's certificate, university graduate, or from the documents that we find at work and we suspect it is fake, or we will fabricate some documents to use it in the test-taking results.
- Evaluate the method by measuring the result performance comparing with other result to ensure our suggested method.

1.4 Importance of the thesis

1. Enhance the performance degree of forgery detection systems.
2. Improve some methods that used in forgery detection systems.
3. Build a dataset for using in researches that related to documents forgeries.

1.5 Scope and limitations:

1.5.1 Scope:

1. Depend on pixels properties in grey level scanned document.
2. We have built our own dataset.

3. Using high-resolution images at least 300x300 pixels, to get clear details for word edges and intensity because we need to use details in our research.

1.5.2 Limitation:

1. Consider only printed and scanned documents. Because almost official documents are printed documents.
2. Deal with only the printed text fabrication.
3. Exception bitmaps images. Because our features depends on some properties that does not exist in the bitmap images.
4. Using high-resolution images at least 300x300 pixels, to get clear details for word edges and intensity because we need to use details in our research.
5. Ignore the copy move and image splicing fabrication because they depend on different techniques to detect.

1.6 Thesis Organization

This thesis has six chapters and it is organized as follows: The second chapter is devoted to concepts of fabrication, and theory background so that the readers will be familiar with the problem of documents fabrication. Chapter three defines the related works and classify them depending on some features, and chapter four presents the proposed methods to detect this fabrication; experiments and the results are covered in the fifth chapter; while the last chapter contains the conclusions and future works.

.

Chapter 2

Theory Background

In this chapter we try to survey for all the background for the digital crimes specially the scanned document fabrication

2.1 Background:

With the highest incidence of computer and internet crimes at these days it became essential to know how to prove or produce a digital evidence that we could use upon the courts or upon any legal entity. So depending on that, the companies and government institutions and non-governmental organizations must take into their minds the possibility of their presence in such issues, they must provide themselves with digital evidence to support their position in front of the legal entity.

When prove the crime the physical evidence became very important for anyone investigating the crime to prove it and to find some guides to the law and the judiciary, this guide needs to research, survey, investigative and it requires scientific research in science and technology. It uses of forensic laboratories that is specialized in public security organs, which seeks to provide expertise to help the discovery of the crimes in a scientific manner.

Forensics "is the application of investigative and analytical techniques that conform to evidentiary standards used in or appropriate for a court of law or other legal context" [2].

Computer Forensics: is the use of specialized techniques for the preservation, identification, extraction, authentication, examination, analysis, interpretation and documentation of digital information. Computer forensics comes into play when a case involves issues relating to the reconstruction of computer system usage, examination of residual data, authentication of data by technical analysis or explanation of technical features of data and computer usage.

"Computer Forensics requires specialized expertise that generally goes beyond normal data collection and preservation techniques available to end-users or system support personnel". [4]

So we can explain that the **Digital forensics** is the using of proven methods scientifically to save, collect, display, identity, analyses, translate, documentation, and validate of digital evidence extracted from digital sources in order to facilitate or promote the building of criminal events, or help to abort any illegal operations [4].

Digital evidence is defined as information and data of value to an investigation that is stored on, received or transmitted by an electronic device [1].

Then we can explain that the Digital evidence is all digital information that may be used as evidence in a case.

The gathering of the digital information may be carried out by confiscation of the storage media (data carrier), the tapping or monitoring of network traffic, or the making of digital copies (forensic images, file copies, etc), of the data held. Although hard copy print outs of digital information are not digital evidence in the strict sense of definition, it is considered a starting point for applying digital evidence gathering in the future [4]

This evidence can be acquired when electronic devices are seized and secured for examination.

Digital evidence [1]:

• Is latent (hidden).

• Crosses jurisdictional borders quickly and easily

• Can be altered damaged or destroyed with little effort

• Can be time sensitive

There are many sources of digital evidence, but we can divide them into three major forensic categories of devices where evidence can be found: Internet based standalone computers or devices, and mobile devices. These areas tend to have different evidence gathering processes, tools and concerns, and different types of crimes tend to lend themselves to one device or the other [1].

We can summarize the stages that needed to gather the digital evidence in four stages as a digital guide that can be supported by the courts and the judicial authorities, and these stages are:

• Gather evidence

• Examine the evidence

•Analysis and review of the evidence

• Make a report of all the digital evidence extracted from evidence

Knowing that the stages described can be relied upon at any type of digital evidence, such as files, changes in operating systems, network data, and other sources of evidence.

As shown in figure (2-1) below we described how to make the data that we get from any data container to good digital evidence that can support us in a court or in internal use to prove the miss use from any employee or user.

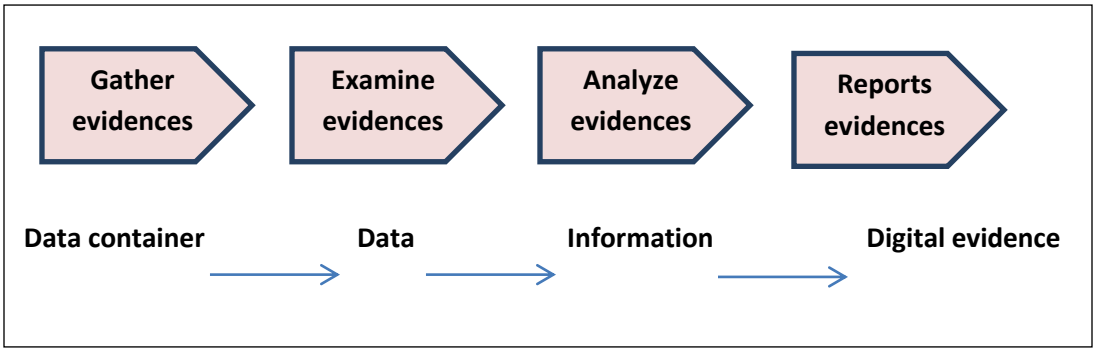


Figure (2.1) Simplify the way the investigation of digital evidence

Data container may be a hard disk, floppy disk, external memory, etc..

One of the most famous digital crimes which use the computer to produce some materials that can be used in an illegal way in the society is the fabricated or forgery document or image

2.2 Concepts of document Fabrication

The crime of forgery appeared and originated with the emergence of writing and the prevalence of use in our life, and writing evolved with the development of civilizations and with increased awareness.

When talking about valuable documents, most people think of passports, ID cards, banknotes, or diplomas. However, it is widely ignored that even every-day's documents like bills and vouchers may be forged to gain financial advantages. we can distinguishes five different document value types [15]:

- Direct value: documents giving access to unconditional and immediate value, e.g. banknotes.

- Indirect value: documents that support a transaction or a right, e.g. diplomas and passports

- Conditional value: documents that give access to a value after the document has been inspected, e.g. admission tickets or cheques.

- Informative value: documents that have no immediate value apart from the information that it contains, e.g. confidential reports

- Fictions value: documents that have no immediate value and that do not contain any valuable information, e.g. stationeries of institutions or companies.

The document fabrication techniques are classified into three general categories [35]:

1. copy/move forgery
2. image splicing
3. image retouching

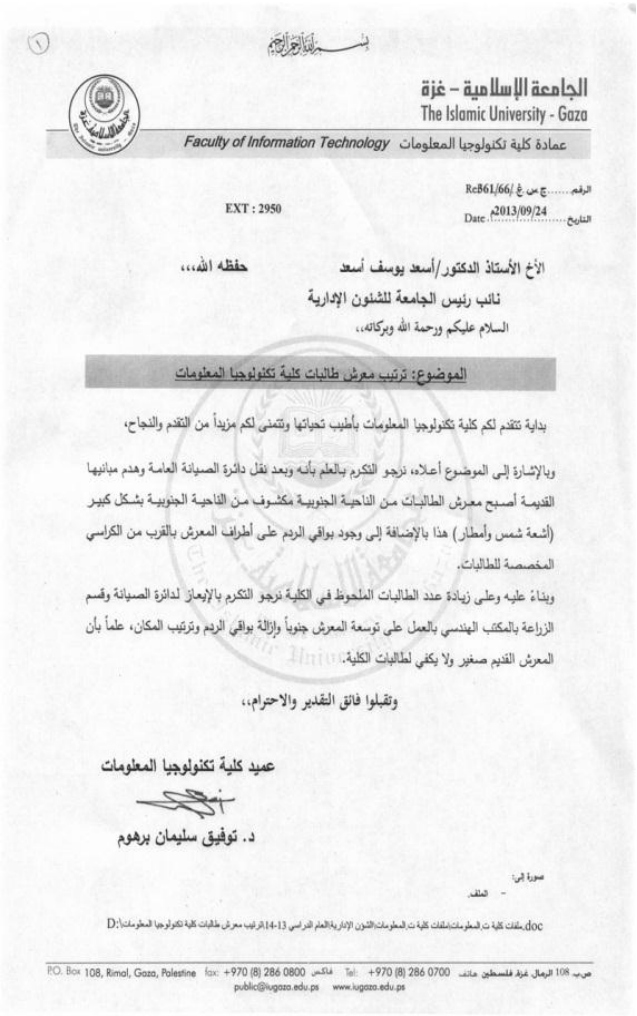
These types of digital forgeries are mainly performed on data, documents, Currency or Cheques, images, and video, but there are some differences between them, so the Copy/Move Forgery is a one of the famous type of image forgery where part of the image is copied and pasted on another part at the same image as shown in figure (2.2).

Figure (2.2) an example for copy/move forgery

The image splicing is almost the same as the copy/move but here the Counterfeiters copy or cut a part of another image and paste it in the target image as shown in figure (2.3)

Figure (2.3) an example for image splicing forgery

The retouching means to make some retouch in the image by changing or adding or deleting some contents of the image as shown in figure (2.4).



بسم الله الرحمن الرحيم

الجامعة الإسلامية – غزة
The Islamic University - Gaza
Faculty of Information Technology عمادة كلية تكنولوجيا المعلومات

الرقم.....ج س غ /Re861/66
التاريخ.....Date 2013/09/24
EXT : 2950

الأخ الأستاذ الدكتور/أسعد يوسف أسعد حفظه الله،،،
نائب رئيس الجامعة للشئون الإدارية
السلام عليكم ورحمة الله وبركاته،،

الموضوع: ترتيب معرض طالبات كلية تكنولوجيا المعلومات

بداية تتقدم لكم كلية تكنولوجيا المعلومات بأطيب تحياتها وتتمنى لكم مزيداً من التقدم والنجاح،

وبالإشارة إلى الموضوع أعلاه، نرجو التكرم بالعلم بأنه وبعد نقل دائرة الصيانة العامة وهدم مبانيها القديمة أصبح معرض الطالبات من الناحية الجنوبية مكشوف من الناحية الجنوبية بشكل كبير (أشعة شمس وأمطار) هذا بالإضافة إلى وجود بواقي الردم على أطراف المعرض بالقرب من الكراسي المخصصة للطالبات.

وبناءً عليه وعلى زيادة عدد الطالبات الملحوظ في الكلية نرجو التكرم بالإيعاز لدائرة الصيانة وقسم الزراعة بالمكتب الهندسي بالعمل على توسعة المعرض جنوباً وإزالة بواقي الردم وترتيب المكان، علماً بأن المعرض القديم صغير ولا يكفي لطالبات الكلية.

وتقبلوا فائق التقدير والاحترام،،

عميد كلية تكنولوجيا المعلومات

د. توفيق سليمان برهوم

صورة إلى:
- الملف.

D:\ملفات كلية ت.المعلومات\ملفات كلية ت.المعلومات\الشئون الإدارية\العام الدراسي 13-14\ترتيب معرض طالبات كلية تكنولوجيا المعلومات.doc

ص.ب 108 الرمال غزة فلسطين هاتف Tel: +970 (8) 286 0700 فاكس fax: +970 (8) 286 0800
P.O. Box 108, Rimal, Gaza, Palestine
public@iugaza.edu.ps www.iugaza.edu.ps

Figure (2.4) an example for image retouching forgery

However, not all these techniques are being used in document forgery. The fabrication of the documents is one of retouching technique that is used in documents forgery. Each kind of manipulation, of which changing the content of any of these, becomes illegal [29] and all of them will produce a false or fabricated document.

. 2.3 Types of document fabrication methods:

There are a lot of methods to fabricate any images or documents. In this section we mentioned to some kinds of these methods which produces the fabricated documents from an original such as:

- Replacing a different photograph in place of photograph of authenticated person.
- Replacing contents in variable regions, through cut or copy-and-paste technique from the same document (Copy Move Fabrication) or from more documents.
- Overlaying new content above actual content.
- Adding new content into existing content.
- Removing some content from existing.
- Changing content by overwriting, intellectually changing character in contents.

2.4 Types of detecting fabrication methods

With systems that trust scanned documents, there is an urgent need to create systems to help in the detection and investigation of fabrication in the documents and in the research about this topic.

Many researches attempts to carry out on original documents instead of scanned documents, like signature verification, detection of forged signature, handwriting forgery, printed data forgery and finding authenticity of printed security documents. In addition, this direction reveals that the above research attempts have been made in the following issues [30] :

1. Discriminating duplicate Cheques from genuine ones using some function.
2. Detecting fabrication or manipulation of printed document by classifying laser and inkjet printouts to compare between their fingerprints.
3. Recognition and verification of bank notes of different country using society of neural networks with addressing on forged bank currencies.

4. Identification of forged handwriting using wrinkles as a feature attempted along with comparison of genuine handwriting.
5. Detect fabrication in document by measuring the spaces between text pixels and Text-Line Alignment.
6. Detect Fabrication by find the similarity between blocks in the image , and that use to detect the copy move fabrication.

Many authors aimed to implement generalized forensic techniques for documents and images based on deterioration of their quality or change in image features. Much work has been done to identify the source of the image such as camera, printer or the scanner [29].

2.5 Summary

In this chapter, we mentioned to general background of the subject, and described some of terms such as (digital forensics, document fabrication and fabricated document), then we listed some methods that use to gather evidence and some evidence sources, and then we talked about the basic concepts of document fabrication. After that we described some techniques that use to fabricate documents, and we also explained some ways that use to detect fabrication in the documents. Then we listed some of the changes that may occur to the documents to make them fabricated documents. And that is our research problem as mentioned at the first chapter.

Chapter 3

Related Works

In this chapter we will discuss some related works for the researchers that try to address the fabrication problem. We try to find the positive and the limitation in their studies.

Many researches related to photos forgery, but only a few of them are related to documents texts forgery and we will classify all of them depending on the methods, techniques and features properties they followed :

3.1 Printers Properties:

Johann, Markus, et. al. [12], produce a system to detect the difference of the edge of the printed character to distinguish between different kinds of printers output. They created a dataset with 1,200 document images from different domains (invoices, contracts and scientific papers) printed by 7 different inkjet and 13 laser printers. Then they recorded the characteristics of each printer to clarify the properties of the edges of the letters and then search for different letters to distinguish them in the document to show the suspected letters.

The whole process they used was divided into two main steps: step one is Feature extraction, and step two is Anomaly detection.

During the feature extraction, they classified the printing technique by classifying the features by examining the documents. In the second step, they analyzed the documents and identified documents, which are not printed with the same printing technique as the majority of the documents.

In addition to the feature extraction, two different unsupervised anomaly detection algorithms - Grubbs and k-NN - have been implement and test. Both show promising results in different test cases. Possible reasons for the varied results in the different test cases have been examined and presented. They created dataset contains unique documents for every used printer. There are three different page layouts,

featuring different difficulties for the feature extraction and anomaly detection process. For every printer a unique dataset has been created, in order to ensure a content independent feature extraction system. Therefore, the goal of their research was to create a system being able to discriminate between different printer types. It should work with unsupervised anomaly detection and documents scanned at a moderately low resolution (400 dpi).

Beusekomet [14] presents a system that uses tracking patterns, integrated into the printing process by many printer manufacturers, to expose the source of a document. Another approach by the same authors is using text-line rotation and alignment to detect documents that have been changed with a malicious intent [8].The process of printer or printing technique recognition has also been studied by several groups.

They generate their data set by collecting documents from the Electronic Frontier Foundation (EFF) to get an overview which printers actually generate tracking dots and which do not. A first sample set of 68 sets of test printouts has been scanned in color using a resolution of 600dpi. Each set consists of 8 pages from the EFF printer test set sheets with this information: Manufacturer of the printer / copier, Serial number of the printer / copier used to generate the print outs, Presence or absence of dots, Manually measured horizontal pattern separation distance, Manually measured vertical pattern separation distance.

The authors presented a method for automatically extracting and classifying the counterfeit protection system codes for color laser printer and copiers. They used 7 bases on the vertical pattern separation (VPS) distance. And this has been extended to also horizontal pattern separation (HPS) distance. So the tracking dots are extracted. Then by using statistics, geometric matching of a search pattern on the translation values between the search pattern and the match can be computed. From these, the HPS and VPS distances can be extracted

3.2 Character Level Features:

Romain, Petra, et.al. [28] Present a method which can automatically detect the fabrication based on some document's features at character level. This method is based on outlier character detection in a discriminate feature space and on the detection of the strictly similar characters.

They compute a feature set for all characters. Then, they classify the character whether fake or original based on a distance between them of the same class .

To build their Dataset they developed software to create synthetic fraudulent document images, each dataset contains 20,000 numerical characters where 5% are fakes.

The alignment error is randomly from 1 to 3pixels (up or down), the skew error is randomly created by modifying the inertia axis of a character by 4 to 8 degrees (clockwise or anticlockwise) and the size error is simulated by increasing or decreasing randomly the size of a character by 5 to 10% as shown an examples for each error in figure (3.1).

Figure (3.1) example of conception errors a) number 6 with different size b) number 4 with different skew and c) the misalignment number 8

The documents were created with the Liberation Serif font at 11pt in a 300 dpi resolution.

They realized three experiments in order to evaluate their method:

1) Shapes similarities/dissimilarities by detection of the fraudulent characters using the Shape comparison, and they distinguish of three cases:

A) The fraudster scans a document, frauds by copying and pasting a set of characters and by emailing it.

B) The fraudster scans a document, frauds by copying and pasting a set of characters and by adding some noise to mask his manipulations and emails it (or send it per mail).

C) The fraudster scans a document, frauds by copying and pasting a set of characters that belong to another document with different font properties and emails it.

2) Outlier detection - imperfection retrieval: the second experiment is related to the detection of the imperfection due to the manipulation of the image by the fraudster.

3) Fraudulent document detection: the last experiment consists of a combination of the two previous ones: copied and pasted characters that are also affected by one or more of the three common imperfections.

Joost, Faisal and Thomas [14], describe an approach for forgery detection using text-line information presented. They suggested that the text-line rotation and alignment can be important clues for detecting tampered documents.

They generated their own data set because there are no public real-world dataset could be found to do a meaningful evaluation on. On the other hand, apart from the observations made during the forgery experiment, no statistics could be found on the methods used by amateur document forgers. So they generate there Data set by using different type of printing such as: Print, Paste and Copy 300 dpi, Two-pass Print Color LaserJet, Two-pass Print LaserJet.

There Originals dataset contained 30 document images that were generated by printing and scanning pages from an electronic document. This dataset is used to learn the distributions of the features for genuine text-lines.

The main idea of their research believes that the text-line skew angle and alignment features have been integrated into a statistical framework for automatically detecting implausible skew angles or alignment distances, and they generate tow lines alignment one of the is the base line alignment, and the other lines in the both sides left and right as shown in figure(3-2) below and they depend to describe if the characters, the whole

word or the line are fabricated or not depend on measuring the irregular skew angels or the distance from the alignment lines .

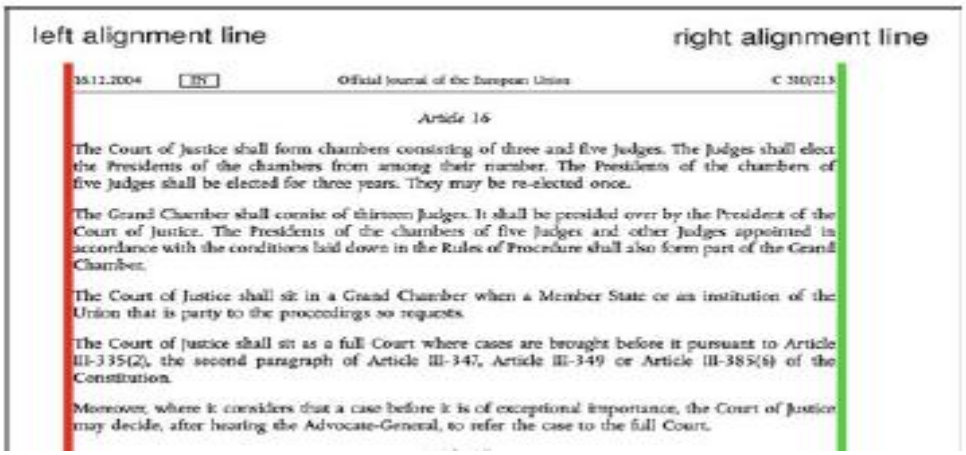


Figure (3.2) an example of base line, left and right alignments

Extensive evaluation of the proposed methods on different datasets has been done to show the usefulness of the approach. We can mention that their accuracy neither depends on the font type nor on the font size.

Lampert et.al. [15] Present a system that uses local features, such as line edge roughness, area difference and correlation coefficients, focusing on single characters of a document. Again, the documents were scanned with a very high resolution (3200 dpi) and a classifier system which needs to be trained has been used.

The dataset they used from generate 26 printouts of 8 laser and 5 inkjet printouts. Then implemented a prototype of the described system in MatLab, and tested it on a dataset. All documents show only text and were scanned at a resolution of 3200 dpi.

Their system uses machine learning to detect different types of printing techniques on the level of individual letters, or even parts of letters as shown in figure (3.3) the different between the edges in the same letters printed in inkjet and laser jet printers.

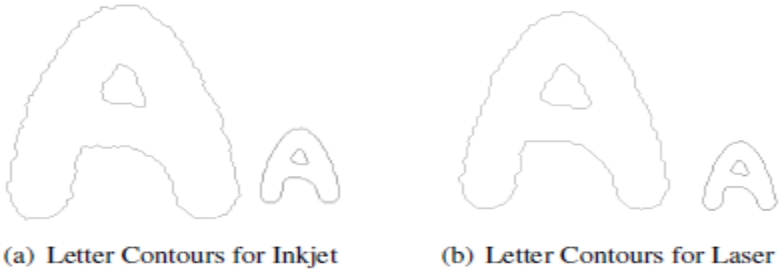


Figure (3.3) the different between edges in inkjet (a) and LaserJet (b)

Furthermore, they have described a setup to detect counterfeit or manipulation of printed documents.

Finally they mentioned that their system is only for assist the user for his own decision. Their classification accuracy average is about 94.8%. However, the classification accuracy varied rather strongly, for some documents reaching 100% but in one case also dropping as low as 78%. We believe that this is caused by lack diversity in the training material

3.3 Page Properties:

Malik [1], described a method to represents an effective and accurate approach to automatic defect detection by try to identifying all kind of defects. Because the non-fabric has regular structure, when breaks the regular structure, therefore, the fabric defects can be detected by monitoring the changes in the paper structure. They describe about 11 kinds of defect of fabrication can be happened, and build a data set images with the same structure of that defect, then they ensure their result by these steps:

1. They developed a fabric defect map, and then from this map, they determined twelve defect types and they used them as the major defects which should be considered during the preprocessing step.

2. They developed simulated plain fabric images either free of defects or with that twelve defect types to understand the behavior of the technique and find the most important factors which effect on their process.

3. Finally they verified the success of their technique in reality, by implemented on real fabric images containing the same defects types we mention above.

3.4 Pixels Properties:

Tiago, Christian et. al. [36] followed a different approach by measuring the color constancy algorithm that depends on specular pixels. In their study, they automatically detect of highly specular regions and isolate it. They propose to segment the image to estimate the illuminant color locally. Recoloring each image region according to its local illuminant estimate yields a so-called illuminant map as shown in figure(3.4) an example of illuminant map for the below picture.

Figure (3.4) example of an illuminate map

They use two datasets. One of them from the images that they captured ourselves, and the second one contains images collected from the internet, the first one contain 200 images 100 forged and 100 original, and the second one contain 50 images with 25 forged and 25 original.

They use five main components for their proposed method:

1) Dense Local Illuminant Estimation (IE): the main goal here is to segment the input image into homogeneous regions. A new image is created where each region is colored with the extracted illuminant color. This result is called illuminant map (IM).

2) Face Extraction: This step may require human interaction. An automated face detector can be employed then the operator sets a bounding box in the image that should be investigated, then crop every bounding box out of each illuminant map, so the remain region will be the illuminant estimates region.

3) Computation of Illuminant Features: for all face regions from the steps above, they computed on the IM values. Each one of them encodes some information for classification.

4) Paired Face Features: For an image with faces, they construct joint feature vectors, consisting of all possible pairs of faces.

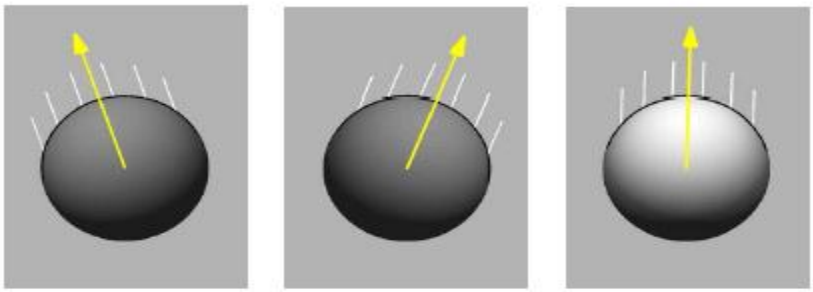
5) Classification: finally they use a machine learning approach to automatically classify the feature vectors. We consider an image as a forgery if at least one pair of faces in the image is classified as inconsistently illuminated. There accuracy was 64.6%.

Farid and Johnson [8] proposed a method, that computes a low-dimensional descriptor of the lighting environment in the image plane. The study estimates the illumination direction from the intensity distribution along manually annotated object boundaries of homogeneous color.

They made their data set by bringing some generated images and natural photographs. The synthetic images consisted of one or more spheres of constant reflectance rendered under either the infinite or local imaging models. The natural photographs were taken outdoors on a clear sunny day (approximating an infinite point

light source), or in a controlled lab setting with a single directional light source (approximating a local point light source).

The main idea in their paper is how to investigate how surfaces of known geometry in the image (plane, sphere, cylinder, etc.) can be used to estimate the third component of the light source direction as shown in figure (3.5)



Figure(3.5) from left to right an example of light sources with single light source for the first and the second spheres positioned at +20 & -20 degree from vertical, and with tow light source positioned at +-20 degree from vertical for the third sphere.

And then their approach will remove the current ambiguity in the light source estimation. They are also investigating a technique to automatically determine which is the best model to describe the underlying image content, is it infinite or local model? So that a forensic analyst does not have to decide which model to use.

Suman, Sharath, and Vasudev [31] Proposed a method to detect the photocopy document fabrication by taking a set of segmented variable regions from a document image as input. And texture features (Gabor/LBP/Edge histogram/Combination) are extracted from them. These features are classified into two classes as fabricated region and non-fabricated region respectively.

The reduced extracted features are queried to the K-means clustering to classify the same as fabricated or not fabricated. The block diagram of the proposed method is shown in figure (3.6)

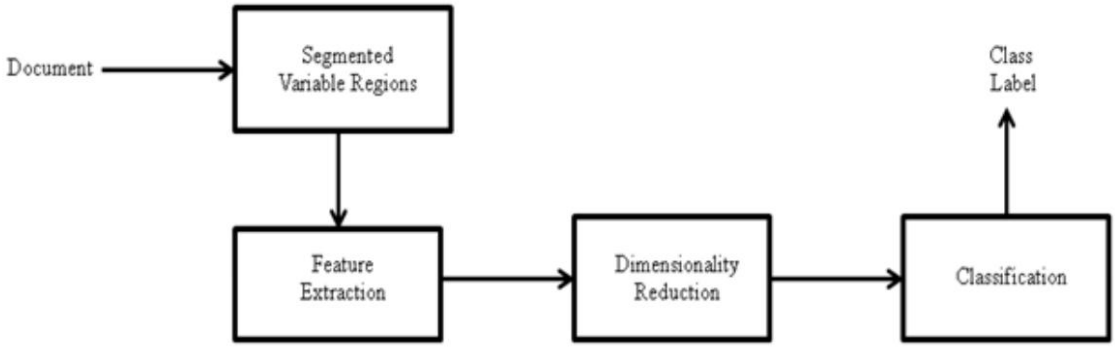


Figure (3.6) Block diagram of stages involved in the proposed method

1. Segmentation of Region of Interest :

   The first step in detection of fabrication is to segment the Region of Interest (ROI).

2. Feature extraction:

   2.1 Gabor filter responses:

   Which is linear filter used for edge detection.

   2.2 Edge Orientation Histogram (EOH):

   The general idea of (EOH), is to represent an image by a histogram obtained from the predominant gradient orientations of its edge pixels.

   2.3 Local Binary Patterns (LBP):

   LBP has been widely used in texture classification because of its simplicity and efficiency. LBP is a simple but efficient operator to describe local image patterns.

3. Dimensionality Reduction:

   Principal component Analysis (PCA) is used for dimensionality reduction. PCA involves a mathematical procedure that transforms a number of possibly correlated variables into a smaller number of uncorrelated variables called principal components. The goal of PCA is to reduce the dimensionality of the data while retaining as much as possible of the variation present in the

original dataset. It is a way of identifying patterns in data, and expressing the data in such a way as to highlight their similarities and differences.

4. Classification:

   K-Means algorithm is an unsupervised classification technique, where the user initiates the algorithm by specifying the number of clusters to be created from feature sets of an image.

   This algorithm splits the given image into different clusters of features in the feature space, each of them defined by its center. Initially each feature in the image is allocated to the nearest cluster.

   Then the new centers are computed with the new clusters. These steps are repeated until convergence. Basically they need to determine the number of clusters K first. Then the centroid will be assumed for these clusters. They could assume random objects as the initial centroids or the first K objects in sequence could also serve as the initial centroids. In the proposed method they consider two clusters because only two classes are required. One is fabrication and another one is non-fabricated.

They pick samples randomly from the database they have made before. Their experimentation is conducted on their database of more than 300 samples. Then they measure the accuracy for each individual features the for the combination between all the features.

The misclassification they achieved is due to dirt and background art in photocopy document. A small amount of fabrication like changing a character or a part of character in the ROI also accounts for misclassification.

3.5 Discussion

Ultimately, all previous researchers proposed verity techniques that concern on forgery detection by extracting different features from the document depends on the printer's finger print, paper, characters or on the pixels properties.

Although, many of these techniques are very promising and innovative, they all have limitations [29], such as using the Arabic language, for the researchers who depend on the character level features and measure the spaces to classify the original and the fabricated characters , and the misclassification when small amount of fabrication like changing a character or a part of character, and also addressing the large-capacity of the image processing software, such as the presence of the ruler that use to locate the place to write on the document accurately aligned with the pattern of writing in the document for whom that use some methods depend on the line alignments and the printers properties.

Therefore, we implement a new method to solve this problem by extracting some useful features from image document depends on the pixels properties. Because our hypothesis relies on the properties of the pixels that pretend the texts which added to the document, comparing with the original text, which exposure to some factors, such as scanning and conversion to other extensions.

And we find difference in the intensity and in the edge gradient of the words characters, and here we have to mention that we also depend on the pixel properties but we cannot compare our method with others because we have different dataset and different environment, but in somehow, we got higher performance than they got in their research.

3.6 Summary

In this chapter we presented some related works and we classified them according to the techniques, methods and features properties that the other researchers depend on such as: printers properties and printers finger print, character level properties, paper properties

and pixels properties, and we described and noticed some of the disadvantages for their studies, such as: using the Arabic language, the fabrication in character or part of character, and the presence of the ruler that use to locate the place to write on the document. We also concluded all the overcome in our research. Finally we mentioned that we cannot compare our method with theirs because of the difference on dataset and environment, but in somehow comparison we got higher performance than the others studies which depend on the pixel properties to detect the retouching fabrication.

Chapter 4

Methodology and Proposed Method

In this chapter we will discuss the proposed method and described all the steps we followed to build it.

4.1 Proposed Method:

The Components of our proposed method are consist of six processes; first three of them are used to make the documents suitable for the next stages, and the other three sub process are included in the stages and in the serial combination as shown in Figure (4.1)

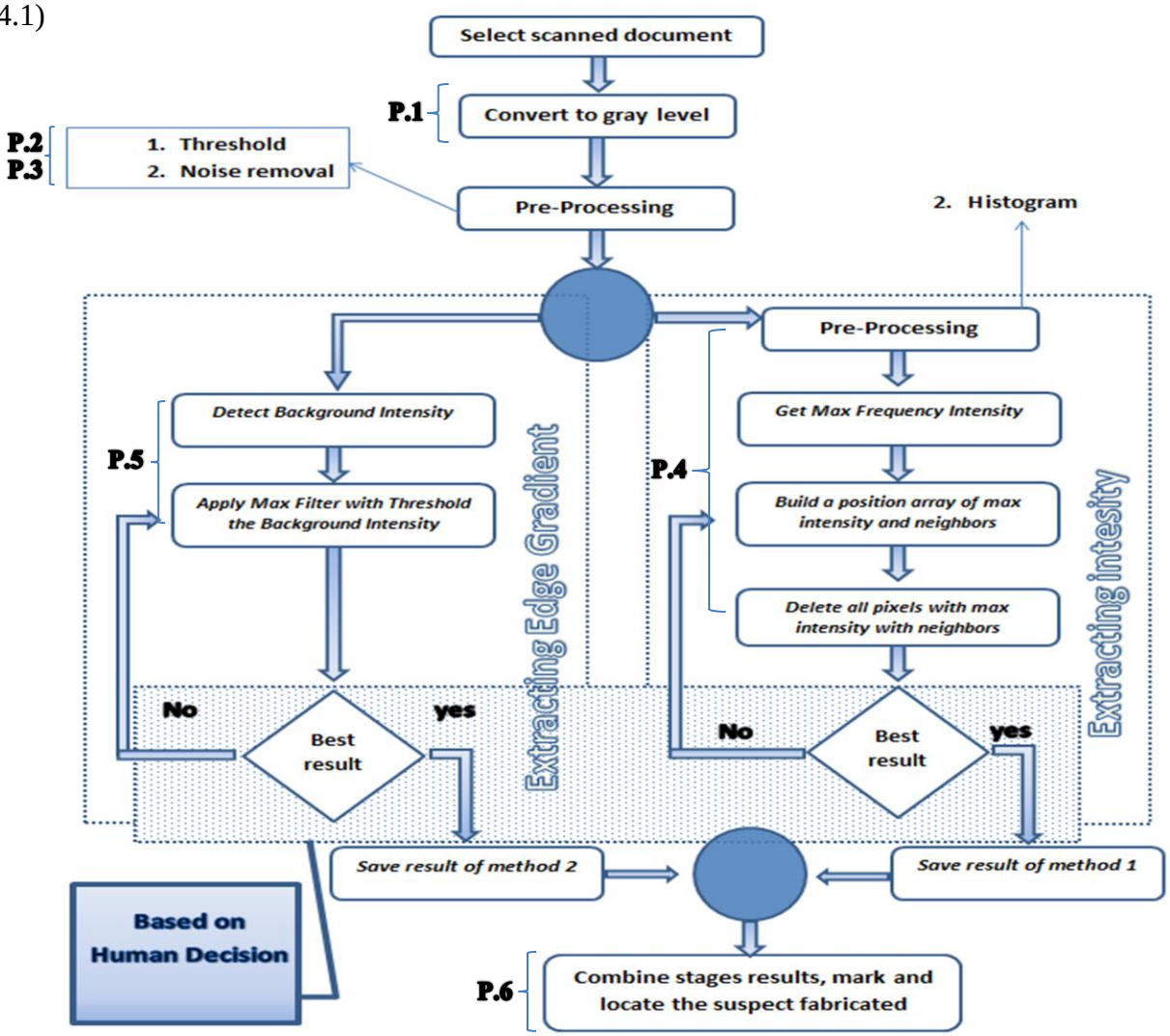


Figure (4.1) the proposed method

The three processes used to make document image suitable for the two stages. The two stages applied sequentially. First stage called Extracting Max Frequency Intensity (EMFI) and second stage called Extracting Edge Gradient (EEG).

4.1.1 Select Scanned Document:

At First, we import an image documents with good resolution at least 300x300 pixels, to get clear details for word edges and intensity. In fact, high resolution helps to get accurate results, but cause a slow extraction results. Also we use a colored or gray scale image and we have convert the color image to gray scale – not black and white – because we depends on pixels intensity to extract some features.

4.1.2 Convert to gray-scale:

The two stages on our method – EMFI and EEG - extract some features that depends on gray scale color level, as well our method import a colored image and gray scale image, so we have to convert the document image to gray scale so that we can extract these features properly.

So, the **first process** we need it on the document image processing in our proposed model is converting to grey level.

To convert a color image to gray scale we have to use some methods depend on the pixels color, If each color pixel is described by a triple (R, G, B) of intensities for red, green, and blue, and there are three algorithms for converting color image to gray scale as describe below [13]:

1. The **lightness** method averages the most prominent and least prominent colors: (max(R, G, B) + min(R, G, B)) / 2.
2. The **average** method simply averages the values: (R + G + B) / 3.
3. The **luminosity** method is a more sophisticated version of the average method. It also averages the values, but it forms a weighted average to account for human

perception. We're more sensitive to green than other colors, so green is weighted most heavily. The formula for luminosity is 0.21 R + 0.72 G + 0.07 B [13].

The example sunflower images in figure (4.2) show the results for each method of converting to gray scale:

(a) Original image

(b) Lightness

(c) Average

(d) Luminosity

Figure (4.2) results for each method of converting to gray scale

We have to mention that we use the Average Method to convert the color document to gray scale because it is the most popular using and also we do not need any changing in the color intensity.

4.1.3 Noise Removal:

The **second process** is noise removal filter. A common problem encountered when scanning documents is ‘noise’ which can occur in an image because paper quality, the typing machine used, or it can be created by scanners during the scanning process [24].

Among other things, noise reduces the accuracy of results tasks of our method such as error in calculate max frequency intensity of pixels in document image.

Normally filters used to remove noise from images. Filters classified into two types, Linear Filters and Non-linear Filters. Linear filters too tend to blur sharp edges, destroy lines and other fine image details, and perform poorly in the presence of signal-dependent noise. With non-linear filters, the noise removed without any attempts to explicitly identify it. The median filter was one of the most popular nonlinear filters for removing Salt & Pepper noise. The noise removed by replacing the window center value by the median value of center neighborhood [19]. So we preferred to use a median filter to remove the noise from the documents.

4.1.4 Threshold:

The third process is Thresholding, and it is probably the most frequently used technique to segment an image. The Thresholding operation is a grey value remapping operation g defined by equation (4.1) :

$$g(v) = \begin{cases} 0 \text{ if } v < t \\ 1 \text{ if } v \geq t \end{cases}$$

Eq. (4.1)

Where v represents a grey value and t is the threshold value. Thresholding maps a grey-valued image to a binary image. After the Thresholding operation, the background of document image will be removing [23].

In fact, more pixels repetition in the document are the background pixels of the document, as shown in figure (4.3). Therefore, we ignored the background pixels by using threshold method.

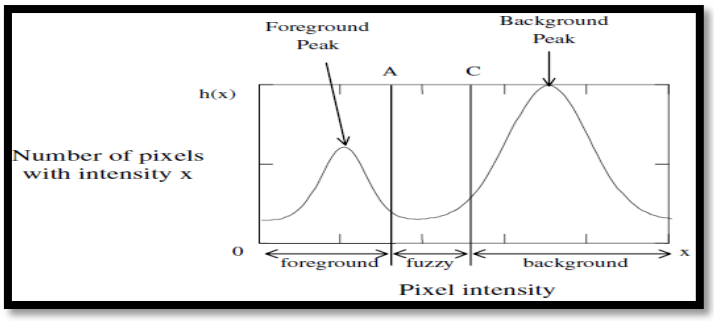


Figure (4.3) Number of pixels and Intensity of background [6]

4.1.5 Stage (1): Extracting Max Frequency Intensity (EMFI):

The visual analysis performed on non-fabricated and fabricated samples collected, exhibit insignificant texture variations in fabricated photocopy document and no such variations are noticed in non-fabricated photocopy document. This led us also to explore the effect on contours of text in photocopy document. Accordingly, a consistent intensity level with smooth edge contour is obtained for non-fabricated text. On the other hand inconsistent intensities with sharp and weak edge contour are resulted for fabricated photocopy text [30].

So, in the most cases the intensity of fabricated words differs from intensity of original word in scanned document, and it is often has solid intensity [30]. Because the fabricated words or letters are made after printing the original document and have another environment such as document processing application and type of paper, specially, each type of paper have internal specifications. Figure (4.4) shows the different levels of gray scale intensity.

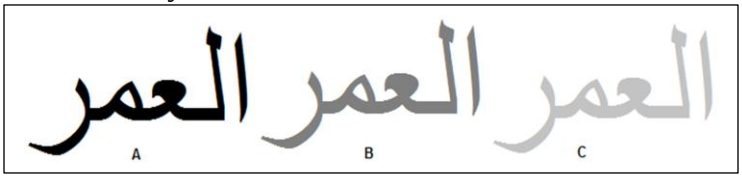


Figure (4.4) Example of gray level of intensity, where intensity of (A) is 0, (B) is 127, and (C) is 195.

In addition, the number of pixels of fabricated words often is less than number of pixels for original words. Therefore, if we isolate the pixels that have a highest frequency of intensity in a document, we can extract the fabricated words or letters in fabricated document. Four steps are using in this stage to extracting words that have a max frequency intensity. These steps were presented in figure (4.1) in page 27.

4.1.5.1 Preprocessing:

Some process will be applied on document image to equip the image for stage one, these processes are histogram and threshold if needed.

4.1.5.1.1 Histogram:

In general, a histogram is the estimation of the probability distribution of a particular type of data. An image histogram is a type of histogram, which offers a graphical representation of the tonal distribution of the gray values in a digital image. By viewing the image's histogram, we can analyze the frequency of appearance of the different gray levels contained in the image. In figure (4.5) we can see an image and its histogram. The histogram shows that the image contains only a fraction of the total range of gray levels. In this case, there are 256 gray levels and the image only has values between approximately "50–100". Therefore, this image has low contrast [27].

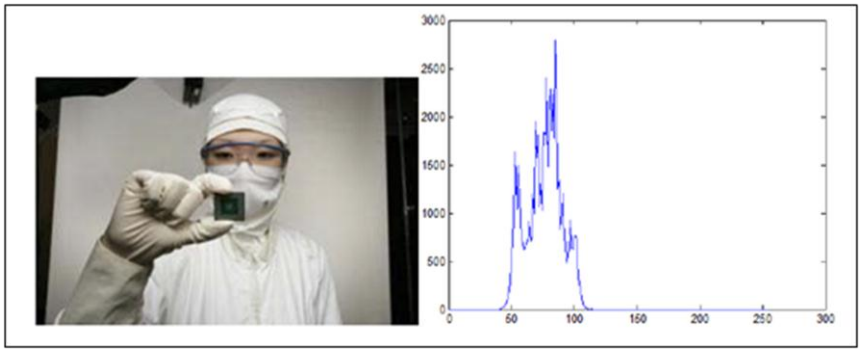


Figure (4.5). An image and its histogram [27].

The result of histogram process for document image is saving in integer array.

4.1.5.2 Get Max Frequency Intensity:

After applying, all previous processes we get an array of grey level for document image and number of iteration for each level in document image; without background intensity.

We extract the max frequency intensity from this array, and build a new array that have the positions of max intensity in document array, and named this array as (MaxIntensityPositionsArray).

4.1.5.3 Delete Max Frequency Intensity Pixels:

The stage 1 depends on user's decision, so, we removed all pixels that have max frequency intensity value, depending on position of pixels that is in MaxIntensityPositionsArray that we had built in previous step to appear the points visually on image of scanned document.

Now, if there are any words – or letters - are appear or disappear visually in image, that is mean these words –or letters- are suspect words. As if did not appear –or disappear- any words or letters visually, we find the next max frequency intensity value from histogram array and build a new MaxIntensityPositionsArray, then delete the pixels of it again. We repeat these steps until appear some words or letters visually.

Figure (4.6) show example of stage one results. Nine words visually appear as suspected words (برهوم , د.توفيق, تكنولوجيا, فائق, أسعد, يوسف, الأستاذ, الأخ, EX:9502). The results are saving as last array of MaxIntensityPositionsArray.

بسم الله الرحمن الرحيم

الجامعة الإسلامية – غزة
The Islamic University - Gaza
Faculty of Information Technology عمادة كلية تكنولوجيا المعلومات

الرقم.....ج س غ/Ref361/66
التاريخ.....Date 2013/09/24م

EXT: 3502

الأخ الأستاذ الدكتور/أسعد يوسف أبو حجر حفظه الله،،،
نائب رئيس الجامعة للشئون الإدارية
السلام عليكم ورحمة الله وبركاته،،

الموضوع: ترتيب معرض طالبات كلية تكنولوجيا المعلومات

بداية تتقدم لكم كلية تكنولوجيا المعلومات بأطيب تحياتها وتتمنى لكم مزيداً من التقدم والنجاح،

وبالإشارة إلى الموضوع أعلاه، نرجو التكرم بالعلم بأنه وبعد نقل دائرة الصيانة العامة وهدم مبانيها القديمة أصبح معرض الطالبات من الناحية الجنوبية مكشوف من الناحية الجنوبية بشكل كبير (أشعة شمس وأمطار) هذا بالإضافة إلى وجود بواقي الردم على أطراف المعرض بالقرب من الكراسي المخصصة للطالبات.

وبناءً عليه وعلى زيادة عدد الطالبات الملحوظ في الكلية نرجو التكرم بالإيعاز لدائرة الصيانة وقسم الزراعة بالمكتب الهندسي بالعمل على توسعة المعرض جنوباً وإزالة بواقي الردم وترتيب المكان، علماً بأن المعرض القديم صغير ولا يكفي لطالبات الكلية.

وتقبلوا فائق التقدير والاحترام،،

عميد كلية تكنولوجيا المعلومات

د. أشرف سليمان ابراهيم

صورة إلى:
- الملف.

D:\ملفات كلية ت.المعلومات\ملفات كلية ت.المعلومات\الشئون الإدارية\العام الدراسي 13-14\ترتيب معرض طالبات كلية تكنولوجيا المعلومات.doc

P.O. Box 108, Rimal, Gaza, Palestine fax: +970 (8) 286 0800 فاكس Tel: +970 (8) 286 0700 ص.ب. 108 الرمال غزة فلسطين هاتف
public@iugaza.edu.ps www.iugaza.edu.ps

Figure (4.6): Results of Stage (1) with one neighbor and two iterations.

To decrease number of iteration for previous steps, we can remove maximum frequency intensity by increasing the number of neighbors.

First neighbor increase over selected intensity by specific value, and the second neighbor decrease over selected intensity by specific value. The specific value depends on expert or users decision too. In practical, two neighbors are enough to detect the suspected word and letters.

4.1.6 Stage (2): Extracting Edge Gradient (EEG):

Accordingly, a consistent intensity level with smooth edge contour is obtained for non-fabricated text. On the other hand inconsistent intensities with sharp and weak edge contour are resulted for fabricated photocopy text [30]. That appeared because there are some stages happened to the original will make it different than the fabricated

text such as the scanning and the converting that happen to the origin text as shown in figure(4.7).

Max filter is use to find the different between the edges of fabricated and non-fabricated text.

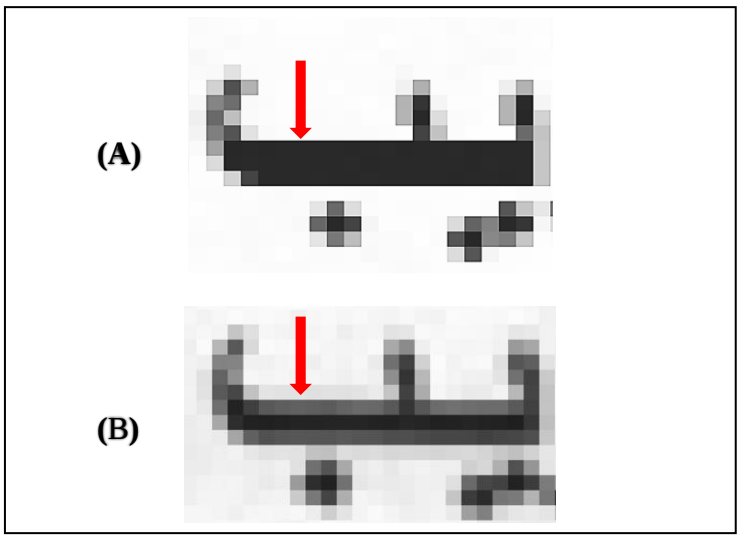


Figure (4.7): (A) Edge Gradient for fabricated text. (B) Edge Gradient for non-fabricated text

4.1.6.1 Max Filter:

The Maximum filter is rank filter in which we select the top ranking grey level from the neighborhood (the maximum grey level) [22]. So we can say that the Maximum filter enhances bright values in the image by increasing its area. For example, given the gray scale 3x3 pixel neighborhood will change to the other one as shown in figure (4.8);

22	77	48
150	77	158
0	77	219

(a)

219	219	219
219	219	219
219	219	219

(b)

Figure (4.8) an example of gray scale 3x3 pixel neighborhood a) before the Max Filter apply and (b) after the Max Filter

The pixels would be changed to 219 as it is the brightest pixel within the current window.

In this stage we will apply a Max filter on image of scanned document. This filter is useful, to make the pixels that have a high intensity spread over all words pixels, that is make the words appear light visually. Usually, applying Max filter on document image one time is enough to appear differs visually, but sometimes we need to repeat this filter more than once to get visually results on image.

All results of this stage appear visually, so the final results of this stage depend on the user's decision. Figure (4.9) show the example of result for stage (2).

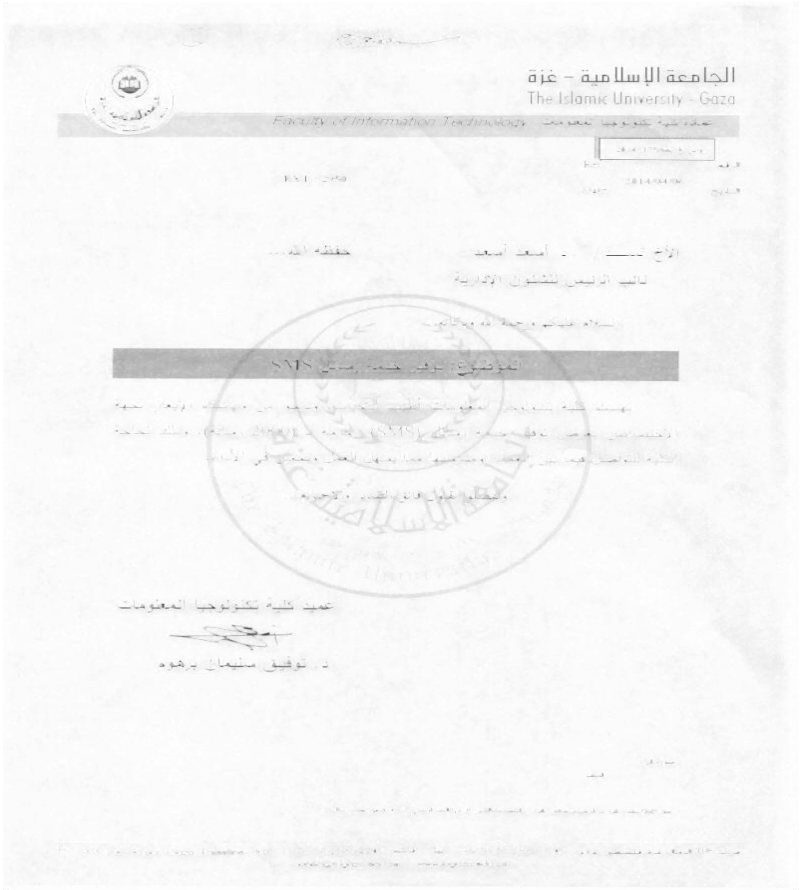
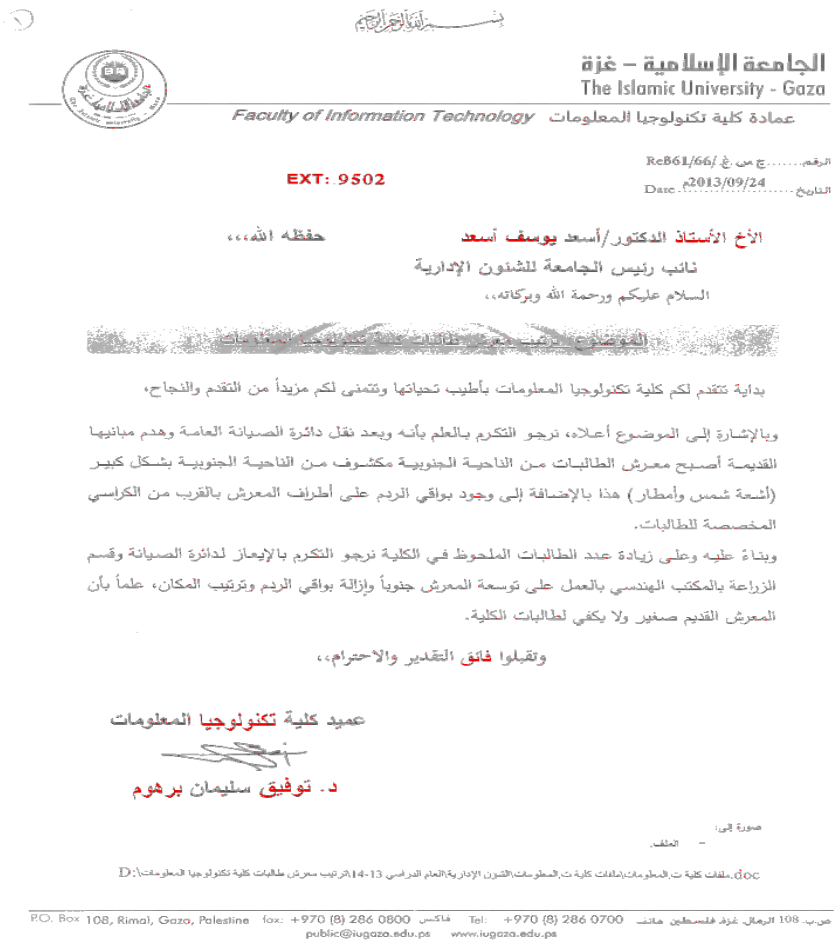


Figure (4.9): Stage (2) result.

4.1.7 Combine Results:

After we applied stage (1) on image of document we get an array of suspected pixels positions for stage (1). In other hand, the result of stage (2) is an image with dark words or letters – that suspected words for stage (2).

In this step, we combined the result of stage (1) and results of stage (2) serially to get the final result of our method, by changing the color of suspected pixel of stage (1) to a red color, and redrawn it on result of stage (2) – dark black pixels -. The joint of two results are the final suspected pixels and appear visually in red color as shown in figure (4.10).



بسم الله الرحمن الرحيم

الجامعة الإسلامية – غزة
The Islamic University - Gaza
عمادة كلية تكنولوجيا المعلومات Faculty of Information Technology

الرقم.....ج س غ/Re861/66
التاريخ.....Date 2013/09/24

EXT: 9502

الأخ الأستاذ الدكتور/أسعد يوسف أسعد
نائب رئيس الجامعة للشئون الإدارية
السلام عليكم ورحمة الله وبركاته،،،

حفظه الله،،،

بداية نتقدم لكم كلية تكنولوجيا المعلومات بأطيب تحياتها ونتمنى لكم مزيداً من التقدم والنجاح،

وبالإشارة إلى الموضوع أعلاه، نرجو التكرم بالعلم بأنه وبعد نقل دائرة الصيانة العامة وهدم مبانيها القديمة أصبح معرش الطالبات من الناحية الجنوبية مكشوف من الناحية الجنوبية بشكل كبير (أشعة شمس وأمطار) هذا بالإضافة إلى وجود بواقي الردم على أطراف المعرش بالقرب من الكراسي المخصصة للطالبات.

وبناءً عليه وعلى زيادة عدد الطالبات الملحوظ في الكلية نرجو التكرم بالإيعاز لدائرة الصيانة وقسم الزراعة بالمكتب الهندسي بالعمل على توسعة المعرش جنوباً وإزالة بواقي الردم وترتيب المكان، علماً بأن المعرش القديم صغير ولا يكفي لطالبات الكلية.

وتقبلوا فائق التقدير والاحترام،،،

عميد كلية تكنولوجيا المعلومات

د. توفيق سليمان برهوم

صورة إلى:
- الملف.

D:\ملفات كلية ت.المعلومات\ملفات كلية ت.المعلومات\الشؤون الإدارية\العام الدراسي 13-14\ترتيب معرش طالبات كلية تكنولوجيا المعلومات.doc

ص.ب. 108 الرمال غزة فلسطين هاتف Tel: +970 (8) 286 0700 فاكس fax: +970 (8) 286 0800 P.O. Box 108, Rimal, Gaza, Palestine
www.iugaza.edu.ps public@iugaza.edu.ps

Figure (4.10): Final results of proposed method.

4.1.8 Summary

In this chapter, we explained our proposed method; where it includes six processes with two stages for detect fabricated words. First three processes are general methods and processes for fabricated document image and make it suitable for the next stages. Process four related to stage one, which include four steps to extract max frequently intensity, called (EMFI). Process five for stage two, which use a Max filter to extracting edge gradient, and called (EEG). Last process, serial combining the results of stage one and stage two, to get final results, that show the suspected fabricated words and letters visually on document image.

Chapter 5

Implementation and Experiments

Implementation

In this chapter we will present how we implement our proposed method. Also, we points to all tools and programming language we used. In addition, we explain all experiments we applied in proposed approach, and discuss all results of these experiments.

5.1 Forgery Detection System (FDS):

Proposed method developed by building a desktop system we named: Forgery Detection System (FDS). The software application of five basic operations as follows:

- Apply some preprocessing.
- Apply Stage (1).
- Apply Stage (2).
- Serial combining the two stages results.
- Save all results.

5.1.1 Implementation:

We implemented (FDS) system by using Microsoft Visual C#.net 2014, because it is powerful language to deal with all methods and libraries for image processing and the system interface is shown in figure (5.1) with some bottoms and fields as describing below:

Figure (5.1): Interface for F.D.S.

Main interface for (FDS) system shown in figure (5.1), and contain the following objects:

- **Main image window**: to show the fabricated document image and show all results for each process or steps for our method.
- **Browse button**: For import, the fabricated document images with many image extensions same as (jpg, bmp, tiff…). Moreover, shown in main window.
- **Process1 button**: to apply stage (1) process on imported image once time and shown in main window.

- **Process1it button**: to apply stage (1) process on imported image many times, and write the number of iteration beside the button, and shown in main window.
- **Number of neighbor's text-box**: type number of neighbors for stage (1) process. The default value is one.
- **Threshold text-box**: type the value of intensity, which used in threshold preprocessing step, to remove background color. Default value is 90.
- **Process2 button and text-box**: to apply stage (2) process on imported image many times. The text box beside the button, use to type how many iteration to apply this process, and shown in main window.
- **Combine button**: combine the result of stage (1) process and result of stage (2) process and shown in main window.
- **Information Status**: show some information and details for the image, and some information for each process.

By selecting the scanned document image to be used for fabrication detection, the system automatically converts the color of image to gray scale.

For stage (1) process, the user must determines the number of neighbors – by default into two neighbors - and threshold of background – by default 120 -. Then apply process1 button – or process1it button -.

The stage (1) will be applied and show the result on main window, and save the result as image in current FDS system path, with name ResultP1.jpg. If we select process1it button, the image result will name as ResultP1-i.jpg, where I is number of iteration.

In addition, the array of result (MaxIntensityPositionsArray) will be also saving in memory.

The user applies stage two after finishing stage one, and determine the background intensity to use it in threshold process – by default 120 -. The FDS system

allows the user to apply this process more than once until we get the best results clear visually. The system saves number of iteration and save the results as image with number of iteration as ResultP2.Ix.jpg, where x is the number of iterations.

Finally, the system serially combines the two results on one image and shows the pixels that were saved in stage one as red points on image of result of stage 2 in main window, and save the image result in current path of FDS system, with name ResultP2P1.jpg.

5.1.2 Applications and Libraries:

We use some applications and libraries to build our proposed method and dataset, here list of these applications:

- **Microsoft Dot Net Framework**: .NET Framework is a software framework includes a large class library known as Framework Class Library (FCL) and provides language interoperability across several programming languages.
- **Microsoft Visual C#:** a powerful language and popular object-oriented language intended for use with .NET.
- **ImageVC Class:** this class made by MS Visual c#, and contains many methods and filters for image processing, and we use it to declare all image objects, and to apply some preprocessing on document image.
- **Adobe Photoshop cc**: most famous and best application for image and photos processing. With powerful filters and tools that use in fabrication. We use it to build our dataset.
- **Microsoft Word 2013**: popular word processing application, and used in write thesis documentation.

5.2 Dataset:

There is no public dataset for document fabrication [15]. Therefore, we built a new dataset by using some original documents from Islamic University of Gaza (IUG) with the following specification:

- Size of document is 800 x 1130 pixels.
- Resolution of document is 300 dpi.
- Compression Mpeg-Layer3 with extension JPG.
- Color level is gray-scale.
- Each document has at least two logos.
- Each document has header and footer.
- Each document has signature.

We used ten types of original documents, each type is fabricated nine times, then the total number of dataset is ninety fabricated documents with 305 fabricated words or letters. Figure (5.2) shows a non-fabricated document, and figure (5.3) shows an example of dataset, a fabricated documents with six fabricated words.

We built our data set by using some experts in multimedia software and they preferred to use Photoshop [40] because of its high potentials in dealing with images and documents, they fabricated the documents in these steps:

1. Scan the documents in color or gray scale with resolution 300dpi at least.
2. Fabricate the documents by using the same fonts in the original documents with some types of filters in Adobe Photoshop, and also without the filters sometimes, and they use several ways such as:
    a. Delete some of the words and replace them with other new words.
    b. Delete some characters from the words and re-write them.
    c. Delete some of the words and re-write without any change.
3. Printout the documents in two kinds of laser jet printers such as HP Laser Jet 1102 and HP Laser Jet 3500.
4. Scan the document again with the same resolution.

بسم الله الرحمن الرحيم

١

الجامعة الإسلامية – غزة
The Islamic University - Gaza
عمادة كلية تكنولوجيا المعلومات Faculty of Information Technology

الرقم.....ج س غ /Re861/66
التاريخ.....2013/09/24م Date

EXT : 2950

الأخ الأستاذ الدكتور/أسعد يوسف أسعد حفظه الله،،،
نائب رئيس الجامعة للشئون الإدارية

السلام عليكم ورحمة الله وبركاته،،

الموضوع: ترتيب معرش طالبات كلية تكنولوجيا المعلومات

بداية تتقدم لكم كلية تكنولوجيا المعلومات بأطيب تحياتها وتتمنى لكم مزيداً من التقدم والنجاح،

وبالإشارة إلى الموضوع أعلاه، نرجو التكرم بالعلم بأنه وبعد نقل دائرة الصيانة العامة وهدم مبانيها القديمة أصبح معرش الطالبات من الناحية الجنوبية مكشوف من الناحية الجنوبية بشكل كبير (أشعة شمس وأمطار) هذا بالإضافة إلى وجود بواقي الردم على أطراف المعرش بالقرب من الكراسي المخصصة للطالبات.

وبناءً عليه وعلى زيادة عدد الطالبات الملحوظ في الكلية نرجو التكرم بالإيعاز لدائرة الصيانة وقسم الزراعة بالمكتب الهندسي بالعمل على توسعة المعرش جنوباً وإزالة بواقي الردم وترتيب المكان، علماً بأن المعرش القديم صغير ولا يكفي لطالبات الكلية.

وتقبلوا فائق التقدير والاحترام،،

عميد كلية تكنولوجيا المعلومات

د. توفيق سليمان برهوم

صورة إلى:

Figure (5.2): non-fabricated document.

بسم الله الرحمن الرحيم

الجامعة الإسلامية – غزة
The Islamic University - Gaza
عمادة كلية تكنولوجيا المعلومات Faculty of Information Technology

الرقم ج س غ /Rc861/66
التاريخ 2013/09/24 Date
EXT: 9502

الأخ الأستاذ الدكتور/أسعد يوسف أسعد حفظه الله،،،
نائب رئيس الجامعة للشئون الإدارية
السلام عليكم ورحمة الله وبركاته،،

الموضوع: ترتيب معرش طالبات كلية تكنولوجيا المعلومات

بداية تتقدم لكم كلية تكنولوجيا المعلومات بأطيب تحياتها وتتمنى لكم مزيداً من التقدم والنجاح،

وبالإشارة إلى الموضوع أعلاه، نرجو التكرم بالعلم بأنه وبعد نقل دائرة الصيانة العامة وهدم مبانيها القديمة أصبح معرش الطالبات من الناحية الجنوبية مكشوف من الناحية الجنوبية بشكل كبير (أشعة شمس وأمطار) هذا بالإضافة إلى وجود بواقي الردم على أطراف المعرش بالقرب من الكراسي المخصصة للطالبات.

وبناءً عليه وعلى زيادة عدد الطالبات الملحوظ في الكلية نرجو التكرم بالإيعاز لدائرة الصيانة وقسم الزراعة بالمكتب الهندسي بالعمل على توسعة المعرش جنوباً وإزالة بواقي الردم وترتيب المكان، علماً بأن المعرش القديم صغير ولا يكفي لطالبات الكلية.

وتقبلوا فائق التقدير والاحترام،،

عميد كلية تكنولوجيا المعلومات

د. توفيق سليمان برهوم

صورة إلى:
- الملف.

D:\ملفات كلية ت.المعلومات\ملفات كلية ت.المعلومات\الشئون الإدارية\العام الدراسي 13-14\ترتيب معرش طالبات كلية تكنولوجيا المعلومات.doc

ص.ب 108 الرمال غزة فلسطين هاتف +970 (8) 286 0700 فاكس +970 (8) 286 0800
P.O. Box 108, Rimal, Gaza, Palestine Tel: +970 (8) 286 0700 Fax: +970 (8) 286 0800
public@iugaza.edu.ps www.iugaza.edu.ps

Figure (5.3): Example of Dataset, a fabricated documents with six fabricated words

5.3 Experiments and Evaluation Results:

In this section, we explain the experiments that applied on proposed method for evaluation and discuss results of each experiment. When we have ten types of documents in our dataset, we need to test each type to evaluate proposed method.

In our experiments, we use the default values for, neighbors and background intensity.

To evaluate our systems we use the confusion matrix depending on the number of words in the document, with recall and precision performance evaluating measures:

The confusion matrix shows how the predictions are made by the model. The rows correspond to the known class of the data, i.e. the labels in the data. The columns correspond to the predictions made by the model. The value of each of element in the matrix is the number of predictions made with the class corresponding to the column for examples with the correct value as represented by the row. Thus, the diagonal elements show the number of correct classifications made for each class, and the off-diagonal elements show the errors made [41].

Precision is a measure of the accuracy provided that a specific class has been predicted. It is defined by:

Precision = tp/(tp + fp)

Recall is a measure of the ability of a prediction model to select instances of a certain class from a data set. It is commonly

also called sensitivity, and corresponds to the true positive rate. It is defined by the formula:

Recall = Sensitivity = tp/(tp+fn). [41]

An example of confusion matrix is shown in table (5.1):

		Dataset	
		Original	**Fab**
Proposed Method	**Original**	True Positive	False Positive
	Fab	True Negative	False negative

Table (5.1) example of confusion matrix table

5.3.1 Document type (1) experiments:

First type of document includes complete Arabic words, English words, dots, dashes, and numbers. And we tested on our proposed method and get the results as shown in table (5.2).

Doc No	Fabric	Forged words or letters	Detect1	Detect2	Detect3
1	1	الأخ	1	0	1
2	2	الأخ – أسعد	2	1	2
3	5	الأخ - أسعد - : - EXT – 9502	5	2	5
4	6	الأخ - الأستاذ - أسعد - : - EXT – 9502	6	3	6
5	7	الأخ - الأستاذ - يوسف - أسعد - : - EXT – 9502	7	4	7
6	8	الأخ - الأستاذ - يوسف - أسعد - فائق - : - EXT – 9502	8	5	8
7	11	الأخ - الأستاذ – يوسف - أسعد - فائق - د - . - توفيق - : - EXT – 9502	11	7	11
8	12	الأخ - الأستاذ - يوسف - أسعد - فائق - د - . - توفيق - برهوم - : - EXT - 9502	12	8	12
9	13	الأخ - الأستاذ - يوسف - أسعد - فائق - د - . - توفيق - برهوم - تكنولوجيا - : - EXT – 9502	13	9	13
Total	65	0	65	39	65

Table (5.2): Results of type (1) experiments.

First Column (Doc No) refers to document number, where each type has nine fabricated document. Second column (Fabric) refers to number of fabricated words or letters on each document of type. Forged words shown the words or letters fabricated in each document. Detect 1 column show number of words or letters detected by stage (1). Detect 2 column show number of words or letters that detect by stage (2). Detect3 column show numbers of words or letters that detect by combine stages.

For Document type 1 we found that, Stage (1) recall was 96.90 and precision was 100, Stage (2) recall was 98.09 and precision was 98.72. The final result for the

performance of serial combining between stage 1 and stage 2 was 96.90 for recall and 100 for precision.

Moreover, the number of iterations for stage (1) and stage (2) to get the best performance is show in table (5.3).

Doc No	St1.It	St2.It
1	5	1
2	5	1
3	1	1
4	1	1
5	1	1
6	1	1
7	1	1
8	1	1
9	1	1

Table (5.3): Number of iteration of stage 1 and stage 2 for type 1.

5.3.2 Document type (2) experiments:

This type includes same previous fabricated objects but we add commas and more dots. The results of experiments for documents type two are show in table (5.4):

Doc No	Fabric	Forged words or letters	Detect1	Detect2	Detect3
1	1	السيد	1	1	1
3	4	السيد ـ أ ـ . ـ د	4	2	4
3	5	السيد ـ أ ـ د ـ . ـ أحمد	5	3	5
4	8	السيد ـ أ ـ د ـ . ـ أحمد ـ حفظه ـ الله ـ ،،،	8	5	8
5	9	السيد ـ أ ـ د ـ . ـ أحمد ـ حفظه ـ الله ـ ،،، ـ الرئيس	9	6	9
6	10	السيد ـ أ ـ د ـ . ـ أحمد ـ حفظه ـ الله ـ ،،، ـ الرئيس ـ نائب	10	7	10
7	11	السيد ـ أ ـ د ـ . ـ أحمد ـ حفظه ـ الله ـ ،،، ـ الرئيس ـ نائب ـ Date	11	8	11
8	12	السيد ـ أ ـ د ـ . ـ أحمد ـ حفظه ـ الله ـ ،،، ـ الرئيس ـ نائب ـ Date ـ عميد	12	9	12
9	13	السيد ـ أ ـ د ـ . ـ أحمد ـ حفظه ـ الله ـ ،،، ـ الرئيس ـ نائب ـ Date ـ عميد ـ برهوم	13	10	13
Total	73	0	73	51	73

Table (5.4): Results of type (2) experiments.

The results of this document type was almost the same as previous type. Stage (1) recall result was 95.02 and precision was 100, Stage (2) recall was 96.41 and

precision was 98.42. The final result for the performance of serial combining between stage 1 and stage 2 was 95.02 for recall and 100 for precision.

Doc No	St1.It	St2.It
1	5	1
2	5	1
3	1	1
4	1	1
5	1	1
6	1	1
7	1	1
8	1	1
9	1	1

Table (5.5): Number of iteration of stage 1 and stage 2 for type 2.

5.3.3 Document type (3) experiments:

Third type includes complete Arabic words, numbers, and some parts of words. Experiments for type 3 and the results are show in table (5.6).

Doc No	Fabric	Forged words or letters	Detect1	Detect2	Detect3
1	1	أسعد	1	1	1
2	1	توفيق	1	1	1
3	1	الأخ	1	1	1
4	1	ورحمة	1	1	1
5	1	التاريخ	1	1	1
6	1	الدال من كلمة الدكتور	1	1	1
7	1	الله	1	0	1
8	1	2950	1	0	1
9	1	يو من كلمة يوسف	1	1	1
Total	9		9	7	9

Table (5.6): Results of type (3) experiments.

The performance results for document type 3 were 99.56 for Stage (1) recall result and it was 100 for the precision result. Stage (2) recall was 99.66 and precision

was 99.90. The final result for the performance of serial combining between stage 1 and stage 2 was 99.56 for recall and 100 for precision.

In other hand, number of iterations for stage (1) increased, and stage (2) has same number of iterations, as show in table (5.7).

Doc No	St1.It	St2.It
1	5	1
2	**5**	**1**
3	**5**	**1**
4	**5**	**1**
5	**5**	**1**
6	**5**	**1**
7	**5**	**1**
8	**5**	**1**
9	**5**	**1**

Table (5.7): Number of iteration of stage 1 and stage 2 for type 3.

5.3.4 Document type (4) experiments:

Type 4 include only completed words. Experiments and the results are show in table (5.8).

Doc No	Fabric	Forged words or letters	Detect1	Detect2	Detect3
1	**1**	**الجرجاوي**	**1**	**1**	**1**
2	**2**	**اياد - الجرجاوي**	**2**	**2**	**2**
3	**3**	**السيد - اياد - الجرجاوي**	**3**	**3**	**3**
4	**1**	**مدير**	**1**	**1**	**1**
5	**2**	**مدير- فرع**	**2**	**2**	**2**
6	**3**	**مدير - فرع - القدس**	**3**	**3**	**3**
7	**1**	**محمود**	**1**	**1**	**1**
8	**3**	**محمود - عبد - الغفور**	**3**	**3**	**3**
9	**1**	**برهوم**	**1**	**1**	**1**
Total	**14**		**14**	**14**	**14**

Table (5.8): Results of type (4) experiments.

The performance results for document type 4 were 99.16 for Stage (1) recall result and it was 100 for the precision result. Stage (2) recall was 99.16 and precision was 100. The final result for the performance of serial combining between stage 1 and stage 2 was 99.16 for recall and 100 for precision. In other hand, number of iterations for stage (1) decreased for all documents, and stage (2) has same number of iterations, as show in table (5.9).

Doc No	St1.It	St2.It
1	3	1
2	3	1
3	3	1
4	3	1
5	3	1
6	3	1
7	3	1
8	3	1
9	3	1

Table (5.9): Number of iteration of stage 1 and stage 2 for type 4.

5.3.5 Document type (5) experiments:

This type has same number of fabricated words, but we add some fabricated letters and dots. Type 5 experiments and the results are show in table (5.10).

Doc No	Fabric	Forged words or letters	Detect1	Detect2	Detect3
1	1	السيد	1	1	1
2	1	المحترم	1	1	1
3	4	السيد ـ أ ـ . ـ أحمد	4	4	4
4	3	أ ـ . ـ المحترم	3	3	3
5	2	أ ـ المطور	2	2	2
6	2	أ ـ رئيس	2	2	2
7	3	أ ـ رئيس ـ الدولية	3	3	3
8	4	أـ أـ . ـ اياد	4	4	4
9	6	أـ أـ . ـ اياد ـ محمود ـ مشتهى	6	6	6
Total	26		26	26	26

Table (5.10): Results of type (5) experiments.

The results of this document type were almost the same as previous. Stage (1) recall result was 98.93, and it was 100 for the precision result. Stage (2) recall was 98.93 and precision was 100. The final result for the performance of serial combining between stage 1 and stage 2 was 98.93 for recall and 100 for precision. In other hand, number of iterations for stage (1) decreased for all documents, and stage (2) has same number of iterations, as show in table (5.11).

Doc No	St1.It	St2.It
1	2	1
2	2	1
3	2	1
4	2	1
5	2	1
6	2	1
7	2	1
8	2	1
9	2	1

Table (5.11): Number of iteration of stage 1 and stage 2 for type 5.

5.3.6 Document type (6) experiments:

Small number of fabricated words, and fabrics some parts of words. Table (5.12) shows all results for type (6) experiments.

Doc No	Fabric	Forged words or letters	Detect1	Detect2	Detect3
1	1	الأخ	1	1	1
2	2	م - .	2	2	2
3	1	الياء من كلمة يامن	1	1	1
4	1	من من كلمة يامن	1	1	1
5	1	مطر	1	1	1
6	1	مسير	1	1	1
7	1	الحكومي	1	1	1
8	1	أعمال	2	1	1
9	1	الفاء من كلمة الفاضل	1	0	1
Total	10		11	9	10

Table (5.12): Results of type (6) experiments.

From table (5.12) we found that, number of detected fabricated words – or part of words - from stage (1) is larger than actually number of fabricated words – or parts -. That is True negative results or positive errors.

The results of this document type were 99.59 for stage (1) recall result, and it was 100 for the precision result. Stage (2) recall was 99.59 and precision was 100. The final result for the performance of serial combining between stage 1 and stage 2 was 99.59 for recall and 100 for precision. However, not all detected words by stage (1) are same words that detected by stage (2). Therefore, Number of iterations for stage (1) and stage (2) shown in table (5.13):

Doc No	St1.It	St2.It
1	3	1
2	3	1
3	3	1
4	3	1
5	3	1
6	3	1
7	3	1
8	3	1
9	3	1

Table (5.13): Number of iteration of stage 1 and stage 2 for type 6.

5.3.7 Document type (7) experiments:

Small number of fabricated words, and fabrics some parts of words more than previous type. Table (5.14) shows all results for type (7) experiments.

Doc No	Fabric	Forged words or letters	Detect1	Detect2	Detect3
1	1	جية من كلمة الخارجية	1	1	1
2	1	لجنة	1	1	1
3	1	التاء من كلمة التدريب	1	1	1
4	1	التاء من كلمة التدريب	1	1	1
5	1	ريب من كلمة التدريب	1	1	1
6	1	محمود	1	0	1
7	1	مرتجى	1	1	1
8	1	ايهاب	1	0	1
9	1	الرقم	1	1	1
Total	9		9	7	9

Table (5.14): Results of type (7) experiments.

The results of this document type were almost the same as previous type. Stage (1) recall result was 99.63 and precision was 100, Stage (2) recall was 99.71 and precision was 100. The final result for the performance of serial combining between stage 1 and stage 2 was 99.63 for recall and 100 for precision.

. Number of iterations for stage (1) and stage (2) are shown in table (5.15).

Doc No	St1.It	St2.It
1	3	1
2	3	1
3	3	1
4	3	1
5	3	1
6	3	1
7	3	1
8	3	1
9	3	1

Table (5.15): Number of iteration of stage 1 and stage 2 for type 7.

5.3.8 Document type (8) experiments:

Medium number of fabricated words, and fabrics some parts of words with English words, and some dots and commas. Table (5.16) show all results for type (8) experiments.

Doc No	Fabric	Forged words or letters	Detect1	Detect2	Detect3
1	2	السيد - T من كلمة EXT	2	2	2
2	5	السيد - T من كلمة EXT - د - . - ابراهيم	5	5	5
3	1	برهوم	1	1	1
4	1	غزة	1	0	1
5	1	عميد	1	0	1
6	1	في	1	0	1
7	1	جامعة	1	1	1
8	2	الله - ،،،	2	2	2
9	1	حفظه	3	1	1
Total	15		17	12	15

Table (5.16): Results of type (8) experiments.

The number of detected fabricated words from stage (1) larger than actually number of fabricated words – or parts -. That is true negative results or positive errors for stage (1), The results of this document type were 99.23 for stage (1) recall result, and it was 100 for the precision result. Stage (2) recall was 99.39 and precision was 99.84. The final result for the performance of serial combining between stage 1 and stage 2 was 99.23 for recall and 100 for precision. However, not all detected words by stage (1) are same words that detected by stage (2). Number of iterations for stage (1) and stage (2) is shown in table (5.17).

Doc No	St1.It	St2.It
1	3	1
2	1	1
3	1	1
4	1	1
5	1	1
6	1	1
7	1	1
8	3	1
9	3	1

Table (5.17): Number of iteration of stage 1 and stage 2 for type 8.

5.3.9 Document type (9) experiments:

In this type, we fabricated large number of fabricated words, without fabrics some parts of words, but with commas, dots, complete words, as shown in table (5.18).

Doc No	Fabric	Forged words or letters	Detect1	Detect2	Detect3
1	1	منصور	1	1	1
2	2	ابراهيم – سعد	2	2	2
3	4	ابراهيم - حفظه - الله - ،،،	3	2	2
4	6	حفظه - الله - ،،، - رئيس - جامعة - غزة	7	3	5
5	6	حفظه - الله - ،،، - رئيس - معهد - غزة	7	3	6
6	1	رئيس	2	1	1
7	3	د - . – سميح	3	2	3
8	4	د - . - سميح – حسنين	4	3	4
9	6	د - . - سميح - حسنين - أبو - الكاس	6	5	6
Total	33		35	22	30

Table (5.18): Results of type (9) experiments.

The number of detected fabricated words from stage (1) is larger than actually number of fabricated words. That is also true negative results or positive errors for stage (1). The results of this document type were 98.30 for stage (1) recall result, and it was 100 for the precision result. Stage (2) recall was 98.86 and precision was 99.43. The final result for the performance of serial combining between stage 1 and stage 2 was 98.46 for recall and 100 for precision. Number of iterations for stage (1) decreased and still stage (2) as same previous types shown in table (5.19).

Doc No	St1.It	St2.It
1	1	1
2	1	1
3	1	1
4	1	1
5	1	1
6	1	1
7	1	1
8	1	1
9	1	1

Table (5.19): Number of iteration of stage 1 and stage 2 for type 9.

5.3.10 Document type (10) experiments:

In this type, we fabricated large number of fabricated words, with all type of fabrics, complete words, parts of words, commas, dots, English words, part of English words and Arabic letters, as shown in table (5.20).

Doc No	Fabric	Forged words or letters	Detect1	Detect2	Detect3
1	3	أ - . – سيف	4	2	3
2	3	حفظه - الله - ،،،	4	2	3
3	6	حفظه - الله - ،،، - د - . - رفيق	7	4	5
4	4	حفظه - الله - ،،، - ابراهيم	5	3	3
5	5	د - . - توفيق - سعيد - الهندي	6	4	4
6	7	د - . - توفيق - سعيد - الهندي - المدير - اتحاد - أنظمة	8	5	6
7	8	د - . - توفيق - سعيد - الهندي - المدير - اتحاد - أنظمة - PI من كلمة PITA	9	6	8
8	8	د - . - توفيق - سعيد - الهندي - المدير - اتحاد - أنظمة - PITA	9	6	8
9	4	اتحاد - أنظمة - PI من كلمة PITA - بر من كلمة برهوم	5	4	4
Total	48		57	36	44

Table (5.20): Results of type (10) experiments.

The number of detected fabricated words from stage (1) is larger than actually number of fabricated words. Here also it is a true negative result or we can say positive errors for stage (1), The results of this document type were 97.83 for stage (1) recall result, and it was 100 for the precision result. Stage (2) recall was 98.33 and precision was 99.49. The final result for the performance of serial combining between stage 1 and stage 2 was 97.98 for recall and 100 for precision

Number of iterations for stage (1) decreased and still stage (2) as same previous types shown in table (5.21):

Doc No	St1.It	St2.It
1	3	1
2	3	1
3	3	1
4	3	1
5	3	1
6	3	1
7	3	1
8	3	1
9	3	1

Table (5.21): Number of iteration of stage 1 and stage 2 for type 10.

Finally, we can conclude the ten previous experiments by following table (5.22):

Total No. of forgery	Forgery Detected by Stage 1	Forgery Detected by Stage2	Combine Forgery Detected	Combining recall result	Combining precision result
305	320	228	299	98.61	99.71

Table (5.22): conclusion for all results

And here we have to mention that we calculated the performance for the serial combining by summation the performance result for the recall and precision results in all the ten types from the experiments for all the combination stages and divided the result for each one to ten to get the serial combination performance.

5.3.11 Other experiments:

In these experiments, we try to test and evaluate our proposed method by using some external documents of our dataset, to evaluate the behavior of our proposed method for other environment.

We get two documents from university and school. First document is a graduation certificate for Al-Quds Open University (QOU), as shown in figure (5.4). This document has only three fabricated words in student name field.

Second document is School Certificate, as shown in figure (5.5). And, include five fabricated words in names.

The results of experiments of two documents are show in table (5.23).

Doc name	Doc No	No of Fabric	Forged words or letters	Detect1	Detect2	Detect3	method 1 it	method 2 it
شهادة المدرسة	1	3	فارس - فادي - ناصر	3	3	3	2	1
المصدقة	2	5	فادية - حسنين - شعيب - توفيق – مشتهى	8	5	5	1	1
	Total	8	0	11	8	8		

Table (5.23): results of other experiments.

جامعة القدس المفتوحة

عمادة القبول والتسجيل والإمتحانات

مصدقة

إستناداً الى أنظمة الجامعة، قرر مجلس الجامعة في جلسته

أربعمائة وسبع وعشرين بتاريخ 2014-05-31

منح فادية حسنين شعيب توفيق مشتهى

المولود في غزة عام 1983

درجة البكالوريوس في العلوم الإدارية والاقتصادية

تخصص إدارة الأعمال

معدل 75.08 وتقدير //جيد جداً//

وذلك في نهاية الفصل الثاني من العام الدراسي 2014/2013

د. جمال إبراهيم

عميد القبول والتسجيل والامتحانات

صدرت في فلسطين بتاريخ 2014-06-22

تحت رقم 1132/01600/78

Figure (5.4): graduation certificate for QOU

مدارس الأنروا واليونسكو للاجئي فلسطين

قطاع غزة

دائرة التربية والتعليم

الصف الأول الأساسي

النتائج المدرسية العام الدراسي ٢٠١٠/٢٠١١

الاسم : فارس فادي ناصر المدرسة : أ/ جباليا الابتدائية المشتركة

الجنسية : فلسطينية الشعبة :

مكان الولادة : جباليا المديرية/القرية/المخيم : محافظة الشمال – مشروع بيت لاهيا – ش. ك

تاريخ الولادة : 2004/7/21 المنطقة : منطقة الشمال

المبحث	النهاية العظمى	النهاية الصغرى	الفصل الأول (50)	الفصل الثاني (50)	المجموع (100)	التقدير	ملاحظات
التربية المهنية	100	50	49	50	99	ممتاز	
اللغة العربية	100	50	50	50	100	ممتاز	
اللغة الإنجليزية	100	50	50	50	98	ممتاز	
الرياضيات	100	50	49	48	97	ممتاز	
العلوم العامة	100	50	50	50	100	ممتاز	
التربية الوطنية والمدنية	100	50	50	49	99	ممتاز	
التربية الفنية	100	50	46	49	95	ممتاز	
المجموع	700	350	344	344	688	ممتاز	
التربية الرياضية	100	50	48	49	97	ممتاز	

النتيجة النهائية : ناجح

ملاحظات مربي/ية الصف : متفوقة أتمنى لها دوام التقدم والرقي

عدد أيام حضور الطالب/ية : (191)

عدد أيام الدوام الرسمي المكتمل : (193)

توقيع مربي/ية الصف : قيسار فريج

ملاحظات ناظر/ة المدرسة : ممتاز

توقيع ناظر/ة المدرسة :

التاريخ : 2011/06/19

UNRWA 21768

1) العلامة والتقدير من 90 – 100 ممتاز 80 – 89 جيد جدا 70 – 79 جيد

60 – 69 متوسط 50 – 59 مقبول 49 فما دون مقصر

2) بدأ الدراسة للعام الدراسي ٢٠١١/٢٠١٢ صباح يوم الأحد الموافق ٢٠١١/٠٩/٠٤م

Figure (5.5): School Certificate.

5.4 Discussions:

Depending on previous experiments, we get many important results as follow:

- There are almost an inverse relationship between number of fabricated words per document, and number of iteration for stage (1).
  When increased number of fabricated word in a single document, we decrease number of iteration for stage (1) to get best results, as found in type (3),(5),(6),(7) and (9) experiments, and shown in figure (5.6).

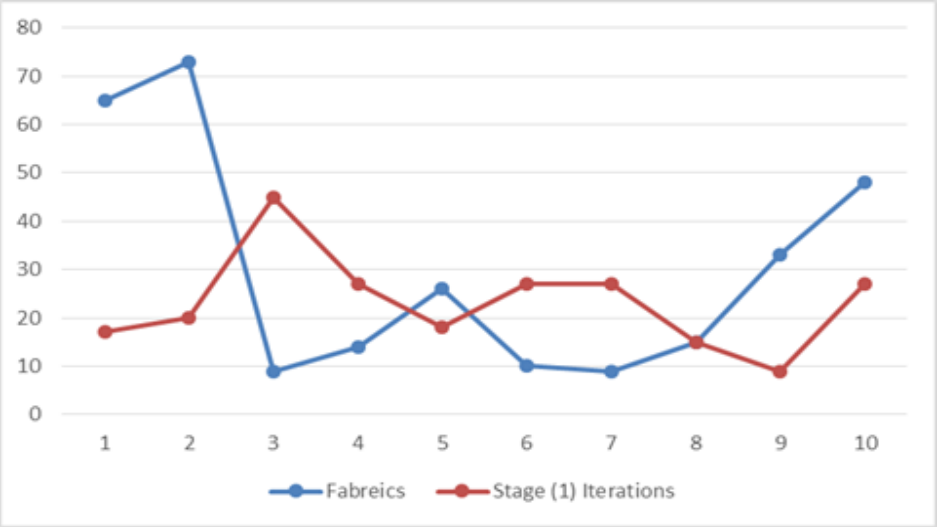


Figure (5.6): Relation between Number of forgeries and Stage (1) iterations.

- There are a direct correlation between number of fabricated words per document, and degree of performance for stage (1) as shown in figure (5.7).
  Whenever increased number of fabricated word in a single document, the performance of of stage (2) will be decrease, as found in type (2) experiments as shown in figure (5.8) for stage (2) performance.

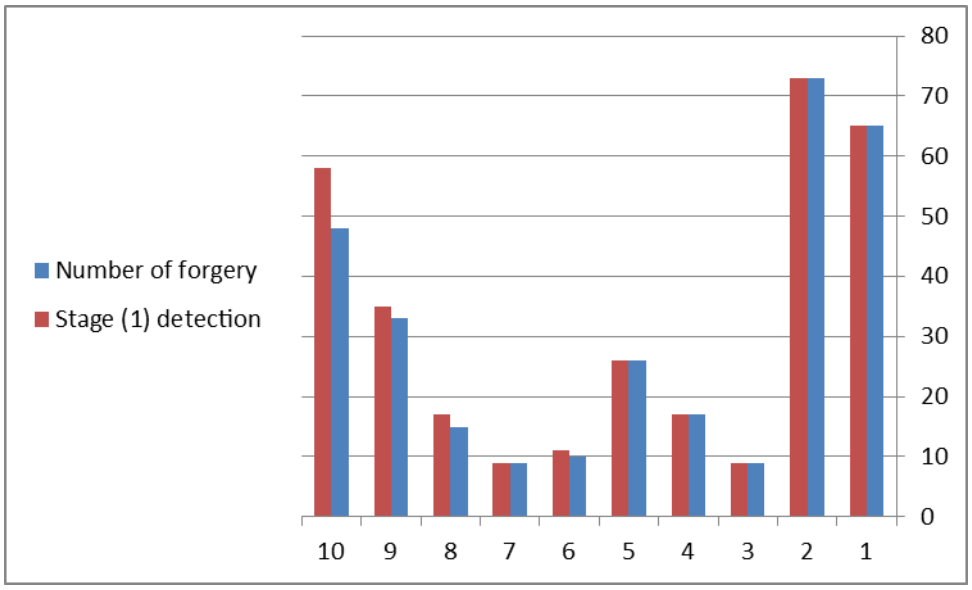


Figure (5.7): stage (1) performance.

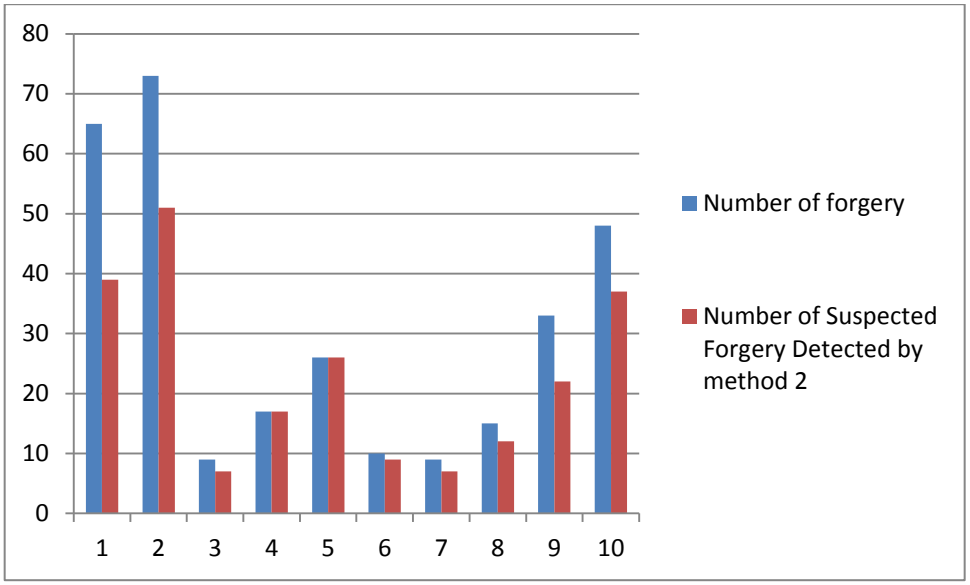
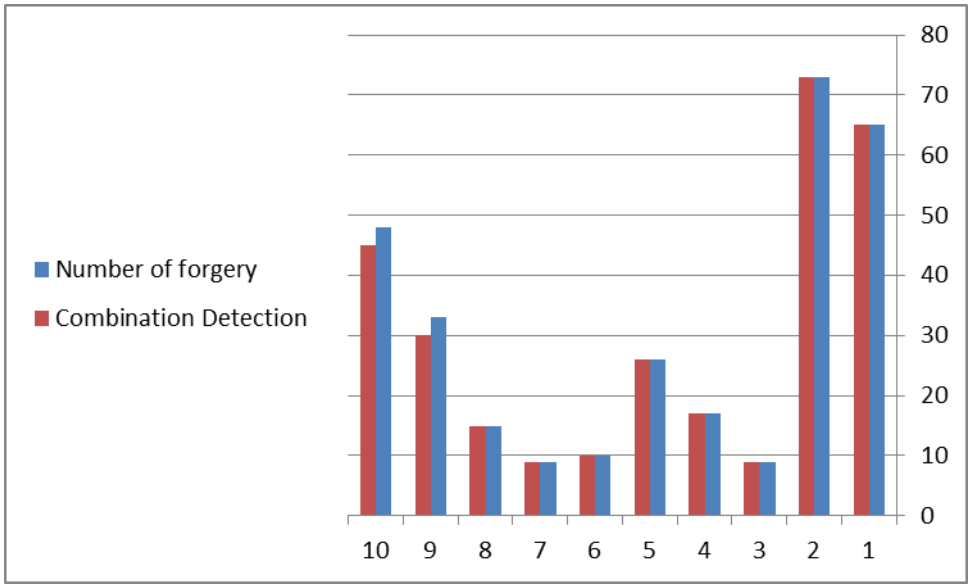


Figure (5.8): Stage (2) performance.

- Number of iterations for stage (1) depends on page type. That means each page – or environment – often has same iteration, regardless of number of fabricated words. This found in all experiments.
- When the fabrication is on parts of word, the degree of performance for stage (2) will be decrease, as found in experiments 3, 6,7,8,9 and 10.

- Total degree of performance by serial combining stage (1) and stage (2) for proposed method is 98.47%. While evaluating by our dataset as shown in figure (5.9).



- Figure (5.9): Serial combination performance.

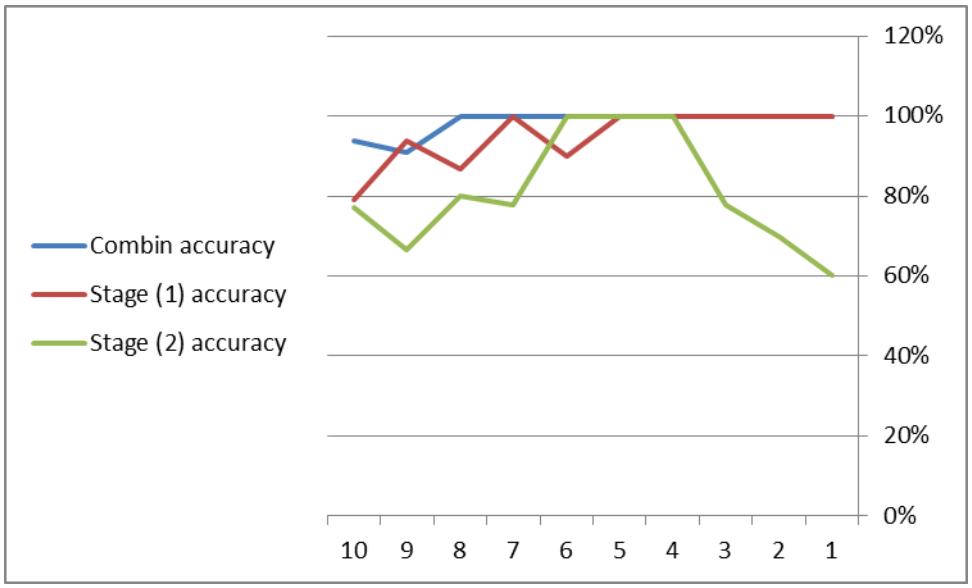


Figure (5.10): All processes performance

-

- Completed words, is best fabricated objects to detect, as found in experiment 4.

- Iterations for stage (2) not considered.
- Neighborhood in stage (1) useful to decreasing number of iterations only, but not effects on results.
- Fabrication on part of words, dots and comma's is the hardest objects for detection.
- The environment of documents is considered in our proposed method.

5.5 Summary

In this chapter, we explicated all experiments that we applied on our proposed method, by using our dataset and some external documents. Each experiment evaluated indecently, and we got many results are discussed including: there are an inverse relationship between number of fabricated words per document and iteration number. To evaluate our method performance we used the confusion matrix depending on the number of words in the document, with recall and precision performance evaluating measures. We used also two external documents to test and evaluate our proposed method on other environment outside our dataset.

Chapter 6

Conclusion and Future work

1. Conclusion

Depending on widely using of the documents, the document fabrication become one of the most famous crime in our life, so to address that crime we reviewed a lot of related works and studied their methods and techniques and classified their techniques depending on some printers, characters, papers and pixels features.

We suggested our new method by extracting some features from the scanned grey level image depending on the pixels properties in the grey level, such as the max frequency intensity of pixels as the first method and the edge gradient as the second one to find some variance between the fabricated text and the original one, then we can detected, located and marked the suspected fabrication by serial combined between each method, and that generated our major new method.

We implemented our forgery detection system then tested and evaluated the results by using 10 types of original documents; each type of document fabricated nine times by some Photoshop experience, so total number of dataset is 90 fabricated documents with 305 fabricated words or letters.

Finally the serial combination performance for the recall result was 98.61for recall result, and it was 99.71 for precision results.

We cannot compare our method to the others method who follow the same method that depend on the Pixels properties because of the difference on dataset and environment, but we in somehow comparison, we achieved higher result performance than the others studies which depend on the pixel properties to detect the retouching fabrication.

2. Future works

For the future work we suggest some ideas that can enhance our Fabrication Detection System, such as:

1. Find some factors to use it in our system to make machine auto decision for the fabricated documents, not only depend on the user experience as it now
2. Try To deal with the stamps and logos in the documents with the text to test all the components in any documents.
3. Extract other features from the document to enhance the performance for the system.
4. Combine some other methods with our method to deal with all kind of image fabrication specially the copy move fabrication.

References

[1] Abdel Salam Malek, "Online Fabric Inspection by Image Processing Technology", Phd Thesis, University of Haute Alsace,(2012).

[2] Chai and Tapert, "Automatic Detection of Handwriting forgery", Proc. 8thInt.Workshop Frontiers Handwriting Recognition(IWFHR-8), Niagara, Canada, pp 264-267, 2002.

[3] Dong ,Wang et.al., "Texture Feature extraction, Pattern Recognition Letter", Vol 6 pp 269-273, 1987.

[4] Efford," Digital Image Processing" , Personal Education Limited, 2000.

[5] "Electronic Crime Scene Investigation", A Guide for First Responders, 2ND) ED, 2008. Department of Justice, Office of Justice Programs, National Institute of Justice. http://www.nij.gov/pubsPsum/ 219941.htm accessed July 5, 2012.

[6] Graham, Saket, et.al., "Separating Text and Background in Degraded Document Images – a Comparison of Global Thresholding Techniques for Multi-Stage Thresholding" , International Workshop on Frontiers in Handwriting Recognition (IWFHR'02), 2002.

[7] Hany Farid, "A Survey of Image Forgery Detection", Dartmouth College, 2008.

[8] Hany Farid and Johnson M. "Exposing Digital Forgeries By Detecting inconsistencies in lighting", ACM Multimedia and security Workshop, 2005.

[9] Hsu, "Image Tampering Detection for Forensic Applications", Phd thesis, Columbia University, 2009.

[10] Huai, Swami, et.al.. "Noise features for image tampering detection". In IEEE International Conference on Image Processing, San Antonio, TX, 2007.

[11] International competition network.org/uploads/library/doc627- "Anti-Cartel Enforcement Manual Cartel Working Group" Subgroup2: Enforcement Techniques ,March, 2010.

[12] Johann, Markus, et.al. , "Document Authentication using Printing Technique Features and Unsupervised Anomaly Detection ". German Research Center for Artificial Intelligence (DFKI GmbH), D-67663 Kaiserslautern, Germany, 2012.

[13] John, "Three algorithms for converting color to gray scale" - August 2009

[14] Joost, Faisal, et.al., "Text-line examination for document forgery detection", IJDAR DOI 10.1007/s10032-011-0181-5, (2012).

[15] Joost, Faisal, et.al., "Document Inspection Using Text-Line Alignment", DAS '10, June 9-11, 2010, Boston, MA, USA, (2010).

[16] Jhon, Soukal, et.al., "Detection of copy move forgery in digital images". In Proceedings of Digital Forensic Research Workshop, August 2003.

[17] Katrin, "Digital Forensics Current and Future Need" Norwegian Information Security Laboraratory (NISLab) Gjovik University College - www.nislab.no, 2010.

[18] Madas, Mohd, et.al., "Off-line signature verification and forgery detection using fuzzy modeling Pattern Recognition" -Vol. 38, pp 341-356, 2005.

[19] MAHESWARI "Noise Removal In Compound Image Using Median Filter" (IJCSE) INTERNATIONAL JOURNAL ON COMPUTER SCIENCE AND ENGINEERING , 1359-1362, VOL. 02, NO. 04, 2010

[20] Mike and Farid. " Detecting photographic composites of people". In 6th International Work shop on Digital Watermarking, Guangzhou, China, 2007.

[21] Murali, Govindraj, et.al. "Comparison and Analysis of Photo Image Forgery Detection Techniques", International Journal on Computational Sciences & Applications (IJCSA) Vo2, No.6, December 2012

[22] Nick Efford "Digital Image processing - a practical introduction using Java " - - chapter 7 - page 180

[23] Prakash Bethapudi Member IEEE , Dr. E. Srinivasa Reddy , Dr.Madhuri.P. "Detection of Malignancy in Digital Mammo grams from Segmented Breast Region Using Morphological Techniques" ,Volume 5, Issue 4 (May. - Jun. 2013), PP 09-12

[24] Proceedings of the International Multi Conference of Engineers and Computer Scientists (MIECS) Hong Kong, Vol I,IMECS 2013, March 13 - 15, 2013

[25] Phillip, Nillius et.al. "Automatic estimation of the projected light source direction". IEEE Conference on Computer Vision and Pattern Recognition, 2001.

[26] Rishi, David, "Preserving boundaries for image texture segmentation using grey level co-occurring probabilities Pattern Recognition" 39 234-245, 2006.

[27] Robert and David, "Histogram Equalization" ,Free scale Semiconductor. Rev. 0, June 2011.

[28] Romain, Petra, et.al., "A System Based On Intrinsic Features for Fraudulent Document Detection". Computer Vision Center, University at Aut`onoma de Barcelona. (2012).

[29] Surbhi and Parvinder, "Document Forgery: The State of Art", International Journal of Research in Engineering and Technology (IJRET) Vol. 2, No. 4, (2013).

[30] Suman, Patgar et.al., "An Unsupervised Intelligent System to Detect Fabrication in Photocopy Document using Geometric Moments and Gray Level Co-Occurrence Matrix", International Journal of Computer Applications Volume 74– No.12 (0975 – 8887), (2013).

[31] Suman, Patgar, et.al., " Detection Of Fabrication In Photocopy Document Using Texture Features Through K-Means Clustering ", Signal & Image Processing : An International Journal (SIPIJ) Vol.5, No.4, August 2014 .

[32] SuanYe, Quansun, et.al., "Detecting digital image forgeries by measuring in consistencies of blocking artifact. In IEEE International Conference on Multimedia and Expo, pages 12–15, Beijing, China, 2007.

[33] Suman, Avcibas, et.al., "Image manipulation detection with binary similarity measures". In European Signal Processing Conference,Turkey, 2005.

[34] Suman, Avcibas, et.al.. "Image manipulation detection". Journal of Electronic Imaging, 15(4):041102, 2006.

[35] Tanzeela Qazi, et.al., "Survey on blind image forgery detection", IET Image Processing, (2013).

[36] Tiago, Christian, et.al., "Exposing Digital Image Forgeries by Illumination Color Classification", IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, VOL. 8, NO. 7,(2013).

[37] Utpal, Biswajith, "Automatic Authenticity Verification of Printed Security Documents", IEEE Computer Society Sixth Indian Conference on Computer vision, Graphics & Image Processing, pp 706-713, 2008.

[38] Vasudev, "Automatic Data Extraction from Pre-Printed Input Data Forms: Some New Approaches", PhD thesis, University of Mysore, India, 2007

[39] Zai and Queiroz, “Identification of bitmap compression history", JPEG detection and quantize estimation. IEEE Transactions on Image Processing, 12(2):230–235, 2003.

[40] http://www.photoshop.com/

[41] http://www.compumine.com/web/public/newsletter/20071/precision-recall

Appendix

Stage 1 code

```
    private void Process1()
        {

            ImageP1.SetImage(ImageDoc.getImage());
            int threshold =
int.Parse(textBoxIgnoreIntensity.Text);
            int[] _maxIntn;
            int _max;
            _maxIntn = ImageP1.HistogramGray();
            int _tmp = 0;
            do
            {
                _max = Array.IndexOf(_maxIntn, _maxIntn.Max());
                _tmp= _maxIntn[_max] ;
                _maxIntn[_max]=0;

            } while (_max > threshold);

            _maxIntn[_max] = _tmp;

            MaxIntensityPos = GetIntensityPositions(_max,
ImageP1.getImage());
            ImageP1.SaveToFile("ResultP1.jpg");
            PictureOut.Image = ImageP1.getImage();
            toolStripStatusLabelMain.Text = "Max Intensity is "
+ _max.ToString() + " = " + _maxIntn.Max().ToString() + "
iterations.    No. Pixels: " + MaxIntensityPos.Count();//+"mini
Intensity is " + _min.ToString() + " = " +
_maxIntn.Min().ToString() + " iterations";
            toolStripStatusLabelProcess.Text = "Process 1
done..";
        }

public List<Point> GetIntensityPositions(int _CurrentIntensity,
Image _img)
        {

            Bitmap sourceImg = new Bitmap(_img);
                List<Point> _IntensityPos = new List<Point>();
                Point _point=new Point();

                for (int i = 0; i < sourceImg.Width; i++)
                    for (int j = 0; j < sourceImg.Height; j++)
```

```
                    {
                        Color Spixel = sourceImg.GetPixel(i, j);
                        if (((Spixel.R + Spixel.G + Spixel.B) /
3) == _CurrentIntensity)
                        {
                            _point.X=i;_point.Y=j;
                            _IntensityPos.Add(_point);
                        }
                    }
                return _IntensityPos;

        }
```

Stage 2 code

```
public Image Max()
        {
            ImageP2.SetImage(ImageDoc.getImage());

            int[] _3x3 = new int[9];
            int _max; int _itns = -1;
            Bitmap sourceImg = new Bitmap(ImageP2.getImage());
            Bitmap TargetImg = new
Bitmap(ImageP2.getImage().Width, ImageP2.getImage().Height);
            Color Spixel;
            int _ind = 0;
            for (int i = 1; i < sourceImg.Height - 1; i++)
                for (int j = 1; j < sourceImg.Width - 1; j++)
                {
                    int _i = 0;
                    for (int v = -1; v < 2; v++)
                        for (int h = -1; h < 2; h++)
                        {
                            Spixel = sourceImg.GetPixel(j + h, i
+ v);
                            _itns = (Spixel.R + Spixel.G +
Spixel.B) / 3;
                            _3x3[_i] = _itns;
                            _i++;
                            _ind++;
                        }

                    if (_3x3[4] == -1)
                        _max = 255;
                    else
                        _max = _3x3.Max();

                    TargetImg.SetPixel(j, i,
Color.FromArgb(_max, _max, _max));
                    _ind = 0;
                }
            return TargetImg;

        }
private void Process2(int ite)
        {
            ImageP2 = new DyImage();
            ImageP2.SetImage(ImageDoc.getImage());
```

```
            if (int.Parse(textBoxP2Iteration.Text) > 0)
            {
                for (int i = 0; i < ite; i++)
                    ImageP2.SetImage(Max());

                PictureOut.Image = ImageP2.getImage();
                ImageP2.SaveToFile("ResultP2.I" +
textBoxP2Iteration.Text + ".jpg");
            }

            toolStripStatusLabelProcess.Text = "Process 2
done..";
        }
```