مجلة ابن الهيثم للعلوم الصرفة والتطبيقية المجلد23 (3) 2010

إقتراح خوارزمية هجينة لإستحداث مفتاح فهرسة لقاعدة البيانات بناءاً على محتويات الصورة

شهلاء طالب عبد الوهاب ،عماد كاظم جبار
قسم علوم الحاسبات، كلية التربية ابن الهيثم، جامعة بغداد
قسم علوم الحاسبات، كلية العلوم، الجامعة التكنلوجية

الخلاصة

يتناول هذا البحث اقتراح طريقة جديدة (HYBRID Algorithm) لحماية الصور والوثائق من التزوير او التغيير الالكتروني وذلك من خلال ضغط البيانات والحصول على مفاتيح استرجاع وتحقق مبنية على محتويات الصورة من قيم رقمية للالوان الاساسية (الاحمر والازرق والاخضر) وخزنها في قاعدة البيانات ومن ثم استخدامها لتدقيق صحة بيانات الوثيقة المطلوبة من خلال إعادة ضغط بياناتها بالطريقة نفسها ومقارنة مفاتيح قيم محتوياتها مع ما مخزون في قاعدة البيانات. إن الطريقة المقترحة في هذا البحث قد تم تطويرها بناء على طرائق سابقة في ضغط واسترجاع البيانات وهما طريقة (Haar Lifting Scheme) وطريقة (D4 Lifting Scheme) وذلك بالدمج بينهما للحصول على خوارزمية هجينة. لقد تم التثبت من دقة وصلاحية الطريقة المقترحة من خلال اجراء اختبارات الضغط والخزن والاسترجاع لبيانات مجموعة من الوثائق المهمة مثل الوثائق الدراسية لمستويات ضغط تصل الى 14 مستوى. كما تمت مقارنة النتائج المستحصلة من الطريقة المقترحة مع مثيلاتها من طرائق الضغط والاسترجاع الاخرى (Haar Lifting Scheme Lifting scheme and D4 كل على حده من حيث دقة اكتشاف التزوير او التغيير ومقدار المساحة الخزنية. أظهرت نتائج الاختبارات ان الطريقة المقترحة تتميز على بقية الطرائق المستخدمة من حيث تحسّسها لابسط تعديل او تغيير في الوثائق التي يراد استرجاعها في حين اظهرت تحسّناً في مقدار المساحة المستخدمة مقارنة بطريقة (D4 Lifting Scheme).

IBN AL- HAITHAM J. FOR PURE & APPL. SCI. VOL.23 (3) 2010

Proposed Hybrid Algorithm for Generate Database Index Key Based on Image Contents

S. T. Abdulwahab, E. K. Jabbar
Department of Computer Science, College of Education Ibn Al-Haitham, University of Baghdad
Department of Computer Science , University of Technology

Abstract

This paper deals with proposing new lifting scheme (HYBRID Algorithm) that is capable of preventing images and documents which are fraud through decomposing there in to the real colors value arrays (red, blue and green) to create retrieval keys for its properties and store it in the database and then check the document originality by retrieve the query image or document through the decomposition described above and compare the predicted color values (retrieval keys) of the query document with those stored in the database. The proposed algorithm has been developed from the two known lifting schemes (Haar and D4) by merging them to find out HYBRID lifting scheme. The validity and accuracy of the proposed algorithm have been evaluated through experiments with the decomposition of database image consists of important documents like college certifications up to maximal decomposition level of 14. The tests results using the HYBRID algorithm were compared with that of the other methods (Haar and D4 Lifting scheme) in terms of the accuracy of discovering forgeries (retrieval accuracy) and the required store memory area. The results illustrate that the HYBRID algorithm show better performance than the others in terms of the sensitivity to any change in the retrieval documents. Also, HYBRID Algorithm exhibits good improvement in terms of the used memory space compared to the results obtained by D4 Lifting scheme.

Key words: *Image Retrieval; Fraud Documents Prevention; HYBRID Algorithm*.

Introduction

Most of database system use images since they show a large amount of information and are easy to store without data entry. The amount of digital images available for the office document is rapidly growing because of that there is a great need for efficient image indexing and access tools. Document which is fraud involves efforts to obtain genuine identity documents through fraudulent means and the alteration of valid documents to be used for fraudulent purposes [1].

In recent years, fraud document has grown increasingly widespread. The sophistication of these schemes has also grown, as document forgers increasingly use computer software and high-resolution digital scanners to play their trade. Additionally, the Internet is being used more frequently to market fake documents to customers. Most losses become the responsibility of the document issuer. Fraud document prevention is an important part of protecting a database since fraud Prevention can occur through a variety of ways [2].

Content-Based Image Retrieval (CBIR) is considered as the process of retrieving desired images or documents from huge databases based on extracted features from the image itself (without indexing a key word). Features are derived directly from the images and they are extracted and analyzed by computer processing [3]. CBIR is a bottleneck of the access for the multimedia databases that deal with text, audio, video and image data which could provide

us with enormous amount of information. Many commercial and research CBIR systems have been built and developed [4]. Content based image retrieval allowing to automatically extract targets according to objective visual contents of image itself like color, texture and shape [5].

In this work, Haar and D4 lifting schemes have been used to decompose color document images into multilevel scale and wavelet coefficients. New lifting scheme has been proposed throughout this paper. The new scheme presents a progressive retrieval strategy, which contributes to flexible compromise between the retrieval accuracy and memory space. The retrieval performances are compared with those of its classical counterpart in terms of retrieval accuracy and memory space.

Lifting Schemes

The lifting scheme is a tool for constructing second-generation wavelets, which are no longer, dilates and translates of one single function. The lifting scheme can be viewed as a process of taking an existing wavelet and modifying it by adding linear combinations of the scaling function at the same level of resolution [6].The wavelet lifting scheme was developed by Wim Sweldens and others. Wavelet lifting scheme algorithms have several advantages like [7]:

1. Allows faster implementation of the wavelet transform.
2. Saves storage by providing an in- place calculation of the wavelet transform.
3. Simplifies determining the inverse wavelet transform.
4. Provides a natural way to introduce and think about wavelets.

Wavelet algorithms are recursive. So the output of one step of the algorithm becomes the input for the next step. The initial input data set consists of $2^{\mathbf{n}}$ elements. Each successive step operates on $2^{\mathbf{n-i}}$ elements, were i = 1 ... n-1. For example, if the initial data set contains 128 elements, the wavelet transform will consist of seven steps on 128, 64, 32, 16, 8, 4, and 2 elements. $step_{j+1}$ follows $step_j$. If element i in step j is being updated, the notation is $step_{j,i}$.

The forward lifting scheme wavelet transform divides the data set being processed into an even half and an odd half [8]. The Lifting Scheme is a method for decomposing wavelet transforms into a set of stages namely: Split, Predict, and Update [9]. A simple lifting scheme forward transform is shown in Fig. (1).

The lifting scheme forward steps illustrated in Figure (1) can be explained as follows:

1. Split Step

This divides the data set into odd and even indexed samples. The original signal is split into even indexed samples, $S_{j,2L}$, and the odd indexed samples, $S_{j,2L+1}$, where j denotes the level of decomposition and L = 0, 1, 2, …, 2^{n-1} are the indices of the elements in the signal. The splitting process is sometimes referred to as the Lazy wavelet transform [10].

2. Predict Step

Predicts the odd elements from the even elements. The linear interpolation function "predicts" that an odd element will be located at the mid-point of a line between its two even neighbors. The difference between the predicted value and the actual value of the odd element replaces the odd element [3].

The odd and even subsets are often highly correlated. Thus, it is possible to predict one from the other. The predict step, where the odd value is "predicted" from the even value is described by the equation 1:

$$odd_{j+1,i} = odd_{j,i} - P\left(even_{j,i}\right) \quad (1)$$

3.Update Step

The update step replaces the even elements with an average. The goal at this step is to maintain some global properties of the original signal in the reduced set [7]. This result in a smoother input for the next step of the next step of the wavelet transforms. The odd elements also represent an approximation of the original data set, which allows filters to be constructed. The update phase follows the predict phase. The original value of the odd elements has been overwritten by the difference between the odd element and its even "predictor" as in equation 1. So in calculating an average the update phase must operate on the differences that are stored in the odd elements [8]:

$$even_{j+1,i} = even_{j,i} - U\left(odd_{j,i}\right) \quad (2)$$

Haar Lifting Scheme

Haar lifting scheme prediction step predicts that the odd element will be equal to the even element. The difference between the predicted value (the even element) and the actual value of the odd element replaces the odd element. For the forward transform iteration *j* and element *i*, the new odd element, *j+1,i* would be [8]:

$$odd_{j+1,i} = odd_{j,i} - even_{j,i} \quad (3)$$

In the Haar lifting scheme, the update step replaces an even element with the average of the even/odd pair. The new odd value is got from old odd value and old even value as in equations 3.

$$even_{j+1,i} = \frac{even_{j,i} + odd_{j,i}}{2} \quad (4)$$

The original value of the $odd_{j,i}$ element has been replaced by the difference between this element and its even predecessor. The new even value is got from old even value and old odd value as in equations 4.

$$odd_{j+1,i} = even_{j,i} + odd_{j+1,i} \quad (5)$$

Substituting this into the average, we get:

$$even_{j+1,i} = \frac{even_{j,i} + even_{j,i} + odd_{j+1,i}}{2} \quad (6)$$

$$even_{j+1,i} = even_{j,i} + \frac{odd_{j+1,i}}{2} \quad (7)$$

The averages (even elements) become the input for the next recursive step of the forward transform. This can be shown in Figure (2).

The number of data elements processed by the wavelet transform must be a power of two. If there are 2^n data elements, the first step of the forward transform will produce 2^{n-1} averages and 2^{n-1} differences (between the prediction and the actual odd element value). These differences are sometimes referred to as wavelet coefficients. Figure (3) shows a 4-steps forward wavelet transform on a 16-element data set.

One of the elegant features of the lifting scheme is that the inverse transform is a mirror of the forward transform as shown in Figure (4). In the case of the Haar transform, additions are substituted for subtractions and subtractions for additions. The merge step replaces the split step.

D4 Lifting Scheme

The Daubechies D4 Lifting Scheme wavelet transforms are composed of Update and Predict steps. In this case a normalization step has been added as well. One forward transform step is shown in Figure (5) [3, 8].

The split step divides the input data into even elements which are stored in the first half of an N element array section (S0 to Shalf-1) and odd elements which are stored in the second half of an N element array section (Shalf to SN-1). In the forward transform equations below the expression S[half+n] references an odd element and S[n] references an even element. Although the diagram above shows two normalization steps, in practice they are folded into a single function.

Forward transform step equations:
Update 1 (U1):
For n = 0 to half-1

$$S[n] = S[n] + \sqrt{3S[half-1]} \tag{8}$$

Where :
S[n]: one dimensional sequence of input samples
Predict (P1):

$$S[half] = S[hafl] - \frac{\sqrt{3}}{4}S[0] - \frac{\sqrt{3}-2}{4}S[half-1] \tag{9}$$

For n = 1 to half-1:

$$S[half+n] = S[hafl+n] - \frac{\sqrt{3}}{4}S[n] - \frac{\sqrt{3}-2}{4}S[n-1] \tag{10}$$

Update 2 (U2):
For n = 0 to half-2

$$S[n] = S[n] - S[half+n+1] \tag{11}$$

$$S[half-1] = S[half-1] - S[half] \tag{12}$$

Normalize:
For n = 0 to half-1

$$S[n] = \frac{\sqrt{3}-1}{\sqrt{2}}S[n] \tag{13}$$

$$S[n+half] = \frac{\sqrt{3}+1}{\sqrt{2}}S[n+half] \tag{14}$$

Proposed Hybrid Lifting Scheme

In archive systems with an amount of digital images that are rapidly growing, there is a great need for efficient image indexing key as access tool in order to fully utilize this massive digital resource, and to retrieve stored data base images. The traditional index key in database like document numbers, document number and date, sequence number, codes and others is just codes to images coding since it does not extract from the image contents. So in this paper we try to create index key from the document image itself [11].

In general, both Haar and D4 algorithms are used mainly for image decomposition. Each one of them has the special features for accuracy, retrieval and memory space required.

In the proposed algorithm, Haar and D4 algorithms have been used as tools for image key generation depends on its features through image multi decomposition until find a numerical value to be the unique image key consists of three parts (red, green and blue color array). Then, this image key has to be stored with the document image in the database in order to use it later on to discover about any changes on the document image stored in the database to prevent fraud.

The proposed HYBRID lifting scheme depends on merging of the lifting schemes (Haar and D4) equations in update steps of D4 in order to find new algorithm with new

features aid for generation image key. Because we did notice that D4 algorithm needs huge memory to perform its steps and to save last numbers and Haar algorithm is not very sensitive to small changes that may be happened in document image. While, our proposed algorithm needs little storage space and very sensitive capacity to small changes.

Therefore, the following update equation will be used:

$$S[n] = S[n] + \{S[half + n]\}/2 \tag{15}$$

The above equation has been taken from Haar algorithm equation (7) with D4 update 1 (U1) terms in equation (8) and replaces the equation (11) of D4 update 2 (U2) with Haar update equation (7).

Figure (6) explains the structure of the proposed HYBRID algorithm U1H which contains the modification of D4 update 1 and U2H contains the modification of D4 update 2. The following steps represent the proposal algorithm index key generation.

Algorithm: HYBRID ALGORITHM FOR GENERATE INDEX KEY BASED ON IMAGE CONTENTS.

Inputs: Database Document Images.

Output: Image Index Key.

Begin

Image scanning;

{Key generation}

For Recursive=1 to 14

For i=1 to n {where n = image width * image height}

Split (image values) {divide the image values to even and odd elements}

Next i

For n =0 to half-1 {U1 for even elements}

$$S[n] = S[n] + \{S[half + n]\}/2$$

Next n

$$S[half] = S[hafl] - \frac{\sqrt{3}}{4} S[0] - \frac{\sqrt{3}-2}{4} S[half - 1]$$ **{For the first value of odd elements only}**

For n =1 to half-1 {P1 for odd elements}

$$S[half + n] = S[hafl + n] - \frac{\sqrt{3}}{4} S[n] - \frac{\sqrt{3}-2}{4} S[n-1]$$

Next n

For n =0 to half-2 {U2 for even elements}

$$S[n] = S[n] + \{S[half + n]\}/2$$

$$S[half - 1] = S[half - 1] - S[half]$$

Next n

For n =0 to half-1 {Normalization}

$$S[n] = (0.517) * S[n]$$

$$S[n + half] = (1.932) * S[half + n]$$

1. Proposed Algorithm Implementation

The implementation of the proposed algorithm is performed on Pentium IV PC with Visual basic compiler and access database software in two stages as follows:
Stage 1: To create index key for input document image.
Stage 2: To discovering fraud document image through retrieval document image.

Stage 1: As show in Figure (7), input new document image and prepare the document image to:

- Compute the unique index key by the proposed algorithm.
- Save the input image and its key index in database.

Stage 2: Check the fraud in documents image when retrieval any document images as show in Figure (8):

- Compute the unique index key to retrieval image.
- Compare this index key with stored index key of the document image.
- If the index key is the same, then the document image is not fraud (original), otherwise, the document image is fraud.
- Display message box as result to upper if statement.

Applications and Results

The experiments focus on discovering forgeries in the document images through the comparison of the retrieval keys between the query document images and the original document images that previously stored in the database. Two important retrieval indicates have been taken into consideration throughout these experiments. The first one is the sensitivity to any change in the query documents image (retrieval accuracy), and the other is the reduction in the memory space required.

In order to verify the validity, accuracy and capability of the proposed HYBRID algorithm for discovering fraud documents, several important document images have been selected for this purpose. These documents include different college certifications. The experimental tests have been carried out in the simplest way so that the fraud document images can be created from the originals after changing the originality through adding small dots or dashes or replacing some marks and names. The detection of the fraud documents will appear through an alarm installed in the program.

The results obtained by using the proposed HYBRID algorithm are shown in Figure (9). It can be noted the capability of the HYBRID algorithm to detect any change in the document originality even if it is very small or negligible.

In the other hand, the same experiments have been repeated but using Haar lifting scheme. The results can be shown in Figure (10). It reveals that there is no detection recorded for the fraud document.

While, when D4 lifting scheme has been used for the same experiments, it detects the fraud document but using high storage memory space for saving index keys compared with HYBRID algorithm as shown in Figure (11).

Table (1) summarizes the results of three types of important document (college certification) using Haar, D4 and the proposed HYBRID algorithms respectively. The table illustrates that the HYBRID algorithm has very good results in terms of fraud detection with small storage space for saving index keys.

Conclusions

Based on the above experiment results, the following conclusion can be drawn:

1. The proposed HYBRID algorithm shows accurate sensitivity against any change in the original documents compared to the other lifting methods.
2. The proposal Hyper algorithm have best performance in terms of discovering fraud documents

IBN AL- HAITHAM J. FOR PURE & APPL. SCI. VOL.23 (3) 2010

3. Harr lifting scheme shows almost no sensitivity against document fraud.
4. D4 lifting scheme exhibits acceptable sensitivity against document fraud but with relatively high memory space required for saving index keys compared Harr and HYBRID algorithm.
5. In terms of the memory space required for saving index keys in the database, the HYBRID algorithm presents less memory space compared to D4 lifting scheme.

References

1. Jim, K.(2008) ,Document Fraud Investigation,(http://www.honestreporting.com/ articles/reports/Document _Fraud_Investigation.asp).

2. Blanco, J. (2002), Guide to Fraud Prevention-Part 4", (http://www.furninfo.com/abso- lutenm/templates /Article_Retailing.asp?articleid=2359&zoneid=7).

3. Latha, Y. M.; Jinaga , B.C. and Reddy, V.S.K. (2008) " A Precise Content-based Color Image Retrieval: Lifting Scheme", ICGST-GVIP Journal, ISSN: 1687-398X , 8, Issue(1): :25-32.

4. Jin, X. and French, J.C. (2003) ,Improving Image Retrieval Effectiveness via Multiple Queries, University of Virginia, Charlottesville, , 86-93.

5. Tody, D. (2003) ,Simple Image Retrieval Interface Concepts and Issues", MMDB'03, New Orleans, Louisiana, USA, 1-6.

6. Pradhan, B.; Sandeep, K.; Mansor, S.; Ramli, A., and Sharif, A.M. (2007), GIS Terrain Data Compression Using Lifting Scheme-A New Dimension, International Journal of The Computer, the Internet and Management, 15, (2): 9 -19.

7. Spires, W. (2005) ,Lossless Image Compression via the Lifting Scheme, University of Central Florida, 22.

8. Kaplan, i. (2003) ,Wavelets and Signal Processing, (http://www.bearcave.com/misl/misl _tech/wavelets/).

9. Islam, S.; Bhuyan, M.S.; Madesa, A.H. and Othman, M. (2007) ,Inverse Discrete Wavelet Transform Processor for JPEG 2000, Proceedings of the International Conference on Electrical Engineering and Informatics Institute Technology Bandung, Indonesia, 240-243.

10. Tashakkori, R.; Tyler, J.M.; Qi, X. and Sallee, C.W. (2008)" Effect of Scanning of Medical Images on the Performance of Lifting", Dept. of Computer Science, Appalachian State University, Boone, NC 28608, 1-10.

11. Rahma, A. and Jabbar, E. K. (2007) "Proposal Database Image Index Key Extracting Algorithm Based on Image Contents(IIKE) ", Journal of Computers, Communication, Control and System Engineering University of Technology 7 (2): 60-67.

IBN AL- HAITHAM J. FOR PURE & APPL. SCI. VOL.23 (3) 2010

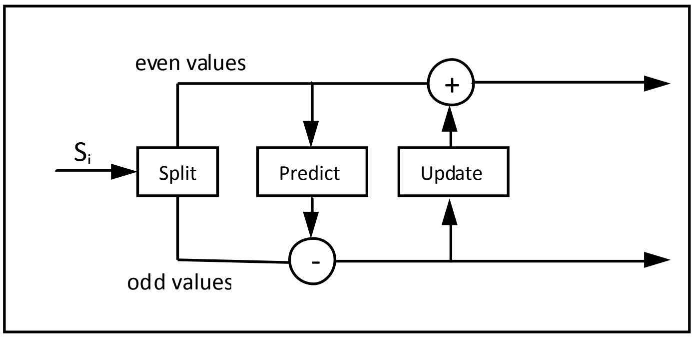


Fig. (1): Lifting scheme forward wavelet transform.

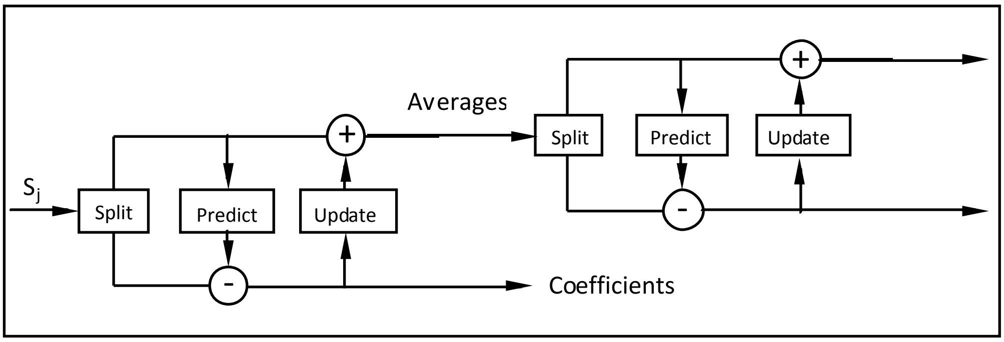


Fig. (2): Two steps in the wavelet lifting scheme Haar forward transform.

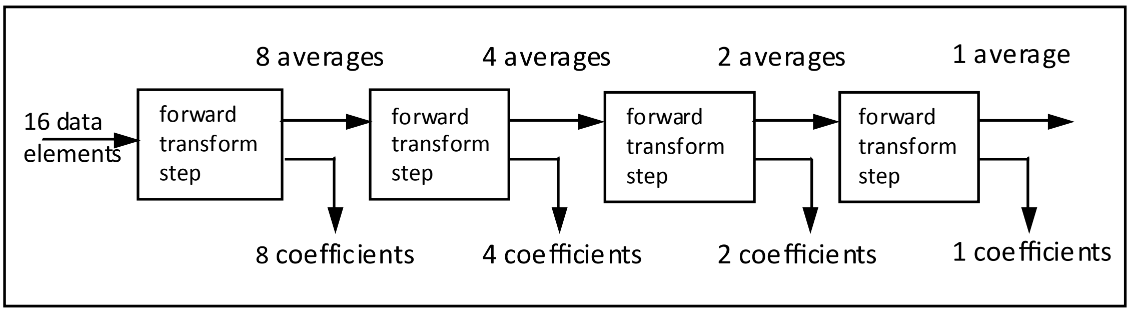


Fig. (3): 4 steps of a 16 element wavelet transform.

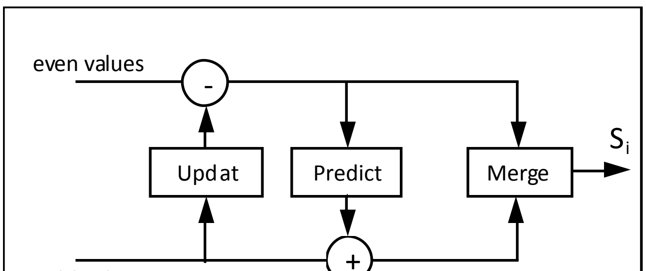


Fig. (4): Lifting scheme inverse transform.

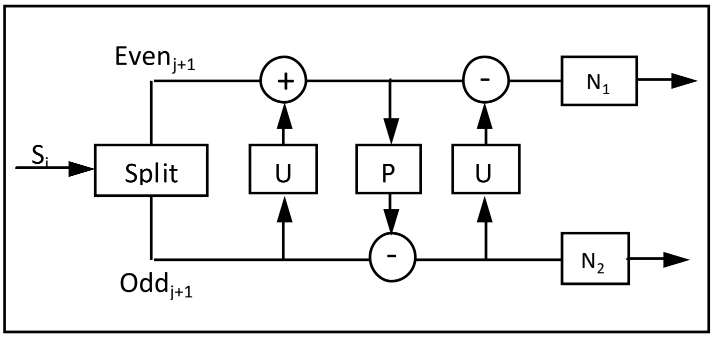


Fig. (5): Forward transform step of the lifting scheme version of the Daubechies D4

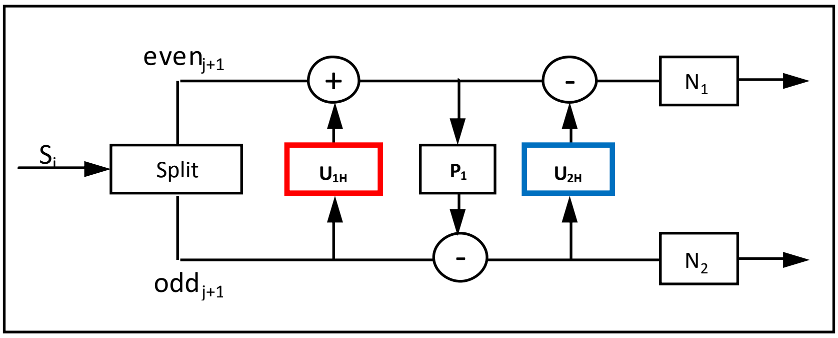


Fig. (6): HYBRID Lifting Scheme

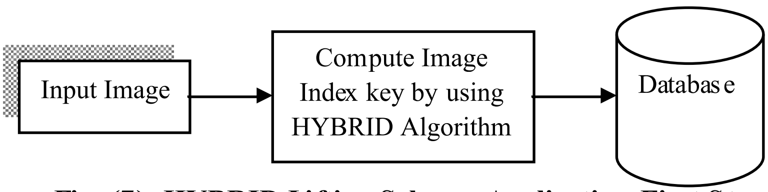


Fig. (7): HYBRID Lifting Scheme Application-First Stage.

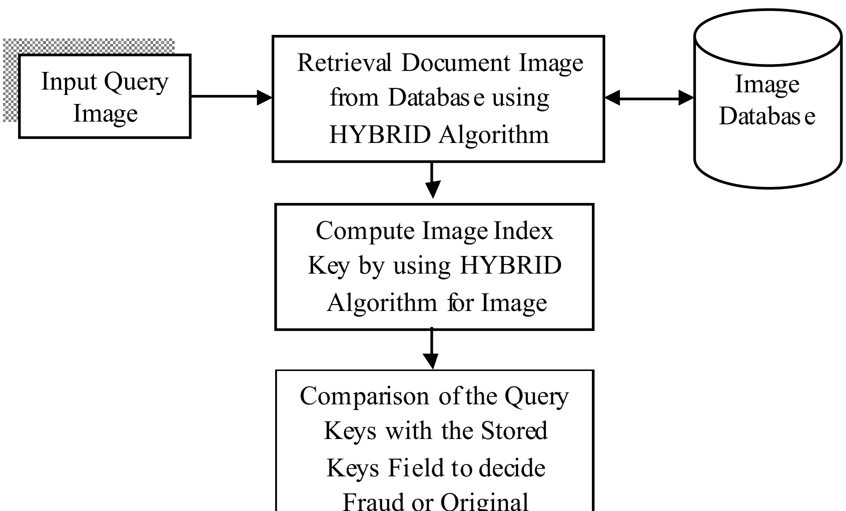
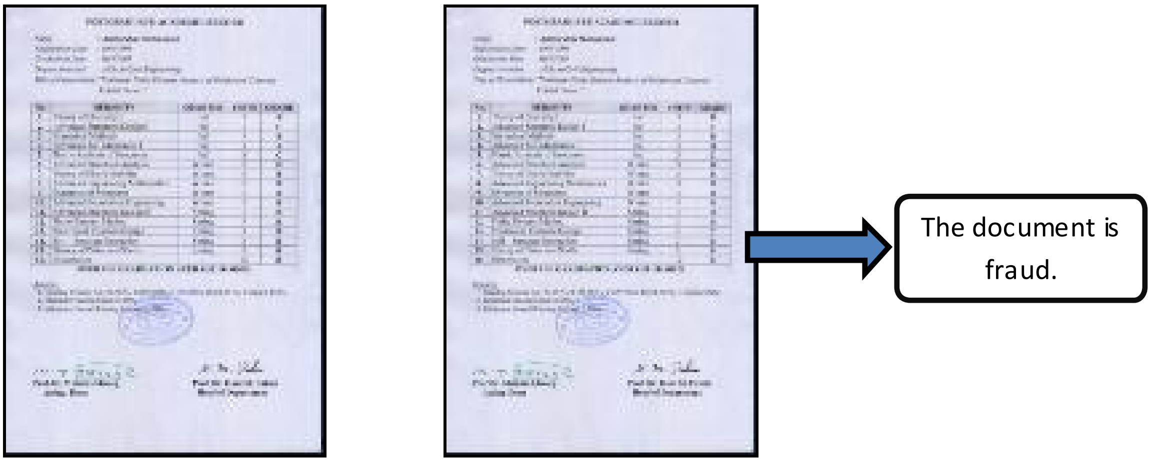
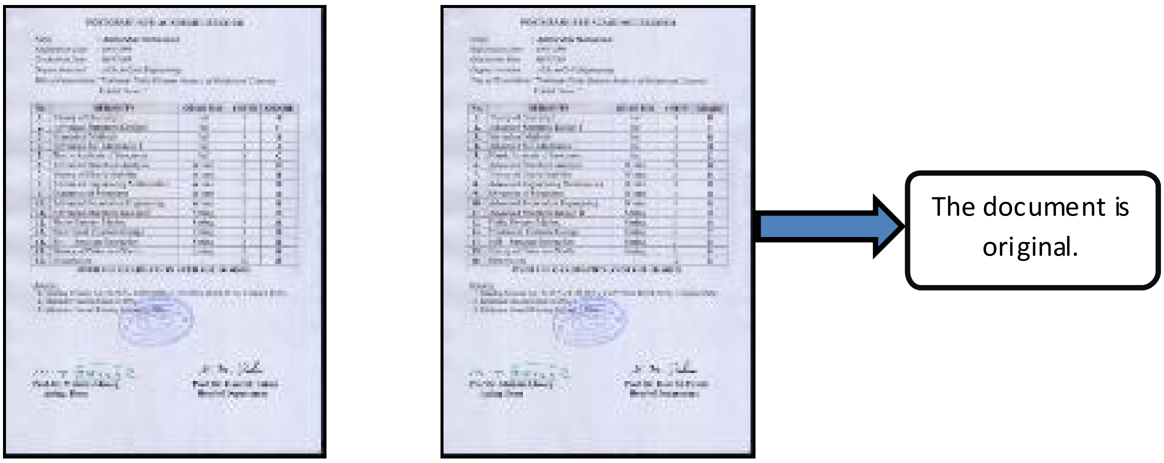


Fig. (8) : HYBRID Lifting Scheme Application-Second Stage .



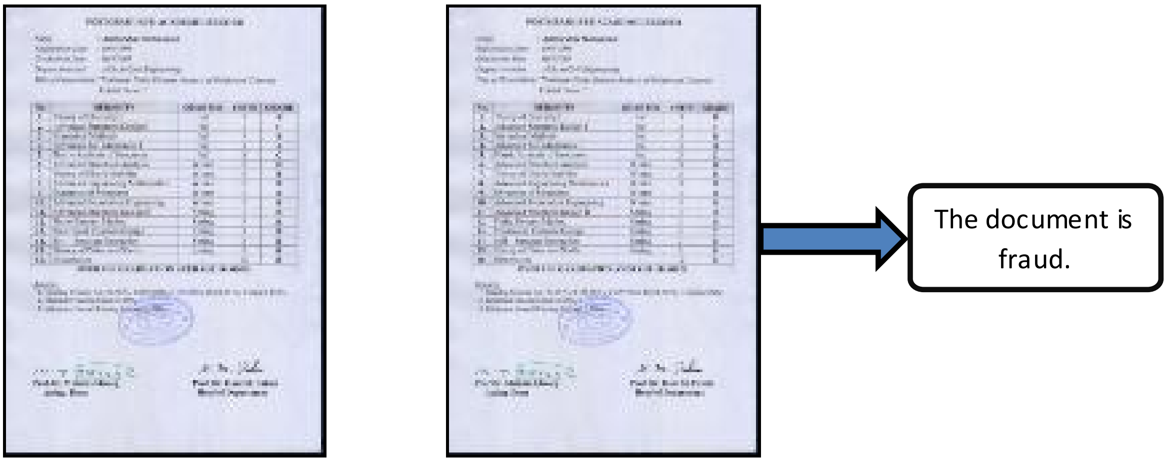
(a) Original Document. (b) Fraud Document (degree no. 4 and 13). (c) Alarm Results.

Fig. (9): Experimental Results for the college certification using HYBRID algorithm.



(a) Original Document. (b) Fraud Document (degree no. 4 and 13). (c) Alarm Results.

Fig. (10): Experiment Results for the college certification using Haar lifting scheme.



(a) Original Document. **(b) Fraud Document (degree no. 4 and 13).** **(c) Alarm Results.**

Fig. (11): Experiment Results for the college certification using D4 lifting scheme.

Table (1): Experiment Results for the college certifications.

Information of Documents

name	Harr red	Haar gree	Haar blue	D4 red	D4 green	D4 blue	Hypered red	Hyperd green	Hyperd blue
ahmed M.Sc.	208	212	227	-28844650736632.6	-20247824033784.2	-27519686070805.9	4912164.639	5004398.772	5366763.681
ahmed M.Sc. f	209	212	227	-28844650736312.8	-28850393548586.1	-28910692798724.4	4912164.627	4963990.811	5508163.455
ahmed B.Sc.	207	210	225	-23918012370033.8	-24321185067134.2	-22593047713345	4969136.707	5061262.850	5423835.678
ahmed B.Sc. f	207	210	225	-23348510380794.1	-23751683677886.2	-22023545724112	4980581.807	5072807.96	5435180.766
ahmed PH.D.	208	211	227	31692031452077.8	35344121963655.3	33864907224144.8	4371339.824	4159741.237	4818289.039
ahmed PH.D. f	208	211	227	-35194054088757.8	-35845847600327.7	-34366329860840	4344633.834	4433035.254	4791383.472
ali B.Sc	208	211	227	25621117811353.9	26172911355908.9	24693393610453.8	5163825.853	5252227.280	5610775.51
ali Ph.D.	208	212	227	-26017503221544.3	-27569206733088.1	-26089778983739	4595013.30	4684314.745	5042862.874

Notes: (f) means fraud document copy.