DCT and Block Based Algorithm for Compound Document Images Segmentation

Ahmed A. Abdelwahab
Dept. of elect, comm. and computer
Faculty of engineering
Helwan university
abdelwahab_99@yahoo.com

Ahmed S. Ali
Dept. of elect, comm. and computer
Faculty of engineering
Helwan university
ah_salah@softhome.net

Mohamed M. Khatan
Dept. of broadcasting engineering
Egypt Media Production City
'EMPC'
eng_mkhatan@yahoo.com

Abstract

Compound images contain a mixture of natural images, text, graphic and background. A scheme to identify those components of the compound image using different features and algorithms for the block-segmentation is proposed in this paper. This scheme is based on a combination of block based algorithms and Discrete Cosine Transform (DCT) based algorithms, which gives more efficient results than using each of them individually. It is also shown that a multiscale context based classification algorithm succeeds when used in conjunction with a well-suited, multidimensional feature space. Experimental results shows that the proposed scheme can reduce the classification error to less than 3% compared to other methods in the literature which gives 12% minimum classification error.

Keywords: *compound images, segmentation, Discrete Cosine Transform based algorithms, block based algorithms, context based classification algorithm.*

1. Introduction

The segmentation of document images into categories such as background, text, and picture, has numerous applications, such as printing processes, compound image compression and pre-processing prior to Optical Character Recognition (OCR).

In general, there are three approaches for image segmentation: object based segmentation, layer based segmentation, and block based segmentation. This paper is focused on block-based segmentation which is very simple and faster than layer based segmentation. The general function for block based segmentation is partitioning an image into non-overlapping (n×n) blocks. This process is based on the information provided by a multidimensional feature space [1, 2].

The previous segmentation algorithms will be discussed in the second section of this paper, while the proposed segmentation technique will be explained in the third section and the experimental results of this proposed algorithm will be shown in the fourth section.

2. Previous segmentation algorithms

Each of these segmentation algorithms computes certain
'activity' score for each block in an image. The block is considered non-text (natural image or background) if its activity score is lower than the assigned threshold, otherwise is considered text [3].

2.1 DCT Based Algorithm

Energy is distributed differently among DCT coefficients for text and non-text blocks. Thus segmentation is achieved by examining the appropriate set of DCT coefficients that capture this difference between text and non-text, and then compare the energy or absolute sum of these Coefficients (the activity) to the threshold. Similarly, the computed absolute sum also shows an important difference between text and non-text, so we do not necessarily have to use energy (absolute sum is faster to compute) [3].

2.2 Wavelet-Based Algorithms

The wavelet transform is an effective means of mapping an image from the space domain to the frequency domain. In a typical wavelet coding scheme, an image's frequency components are subdivided recursively, refining the lower sub band (in the two-band decomposition case) at each step [2, 3].

2.3 Block Based Algorithms

The following algorithms only rely on the n-by-n image blocks, without using any transform [1].

2.3.1 Histogram of Pixel Intensity. This feature based on the histogram of pixel intensity in a block, which is a count of the number of modes, where a mode is defined as a value that is a local maximum and the cumulative probability around it is above a pre-selected threshold. Background blocks have one mode, text blocks have two modes, and picture blocks have multimode.

2.3.2 Mean and Standard Deviation. The mean of pixel intensities (μ) in an image block is a useful feature for segmentation. The three classes, background, natural image, and text can be separated into three bands based on μ.

The standard deviation of pixel intensities (σ) in an image block is a feasible feature for segmentation of scanned images. It can be taken as a primary feature for identifying background

blocks. The standard deviation of pixel intensities (σ) is very small for background images.

2.3.3 Active Pixels. A count of active pixels (α) is defined as the number of pixels whose intensity falls below the threshold ($\mu - k \cdot \sigma$), where k is a constant to be selected. This feature follows the method of adaptive thresholding in the sense that a unique threshold is applied for each block of image.

2.3.4 Sum of Second Derivatives of Average Intensity. Text blocks tend to have many edges, which imply that the average intensity tends to change greatly when traversing the block. The sum of average intensity of second derivative magnitudes (Iav_x , Iav_y) of Savitzky-Golay filter is evaluated at all pixels along the block width. It gives a quantification of edges in the horizontal direction. The summation along the block height similarly provides a quantitative idea about edges in the vertical direction.

2.3.5 Gradient Vector Direction. The gradient vector is easily defined in terms of its horizontal and vertical components. Many gradient vectors point in the four cardinal directions due to the alignment of text edges. Picture blocks and background blocks do not produce such peaks.

2.3.6 Six dimension feature space segmentation algorithm. The simplest block-segmentation algorithms use a single block size which it breaks up the complete image into a grid of blocks and then use a single feature to classify individual blocks. In these algorithms, the choice of block size is critical. A large block enclosing a contiguous region of the same class will tend to have conspicuous features. However, the likelihood of containing multiple classes in a single block increases with increased block size. On the other hand, there is a higher probability that small blocks will contain a single class. Features, however, are less conclusive for segmentation. When many features are used in conjunction to segment image blocks, classification error is reduced. The six dimension feature space segmentation algorithm [1] is using a multiple features which are mean (μ), standard deviation (σ), active pixel (α), sum of second derivatives of average intensity (Dx, Dy),and gradient vector direction (gc). The algorithm is improved by applying context-based classification algorithm [2], which will be discussed in the next subsection.

2.3.7 Multiscale Architecture and Context-based Classification. Multiscale algorithms can take advantage of both the prominent features of large blocks and the likelihood of a single class in small blocks. This type of algorithms begins segmentation with large blocks. If a large block cannot be classified, it is subdivided into four sub-blocks for further evaluation. This multiscale approach resolves an image starting with coarse classification and progressing to fine segmentation. The multiscale technique is further enhanced by using context information. Performance is improved when information gathered during the first pass segmentation of surrounding blocks is used to help classify the unclassified blocks. Rules have been developed to implement such a context-based algorithm. It has been shown that classification error is reduced when a context-based approach supplements the six dimension features segmentation and thus it is a useful component of a block-segmentation algorithm. The main drawback of the six dimension features segmentation algorithm combined with context-based algorithm is that the size of the block is still large. Thus a part of large block may has a class and another part of the same block has another class and the decision will be taken for just one class only from the two classes. The context algorithm depends only on the adjacent blocks. Therefore if the six features algorithm gives a wrong decision for a block and the adjacent block is unclassified, the context algorithm will take a wrong decision, therefore six dimension features combined with context- based algorithm in [1] has a lot of blocks that are unclassified or have a wrong classification so the classification error is still large. In this paper a new efficient segmentation algorithm will be proposed to

avoid the drawbacks of the six dimension features segmentation algorithm combined with context algorithm.
The rest of this paper is organized as follow, in section three the proposed segmentation algorithm is described in details. Section four shows the experimental results for the proposed algorithm and the comparison with some previous algorithms. Finally, the conclusion of the paper is given in section five.

3. The proposed segmentation algorithm

The proposed algorithm is based on a seven features which are the six dimension features (mean (μ), standard deviation (σ), active pixel (α), sum of second derivatives of average intensity (Dx, Dy), gradient vector direction (gc)), and the DCT coefficient absolute sum (dct) feature. The algorithm consists of three stages. The first stage called first pass segmentation. This stage is suitable for classify text regions especially which have large fonts. The next stage called first pass segmentation combined with context-based algorithm. This stage makes a fine classification for the blocks which can not have a classification (unclassified blocks) in the first stage. The last stage is called the second pass segmentation. This stage makes a fine classification for blocks which have a wrong classification in the first stage

3.1 First pass segmentation

In first pass segmentation stage the image is divided into non-overlapping (64*64) blocks. This large size of (64*64) is suitable to classify the text regions especially which have large fonts. In this stage the algorithm uses the seven features as basis for segmentation. Each block has to be classified either as a background, natural image, or text. This happens by comparing an assigned threshold to the calculated results from the features. If the block doesn't meet any of the above classification then it is called "unclassified".
At the beginning the block will be tested if it is background or not, the decision criterion is based on the use of standard deviation and DCT coefficient absolute sum features. Classification will be achieved by comparing them to the calculated thresholds to make a suitable decision for (64 * 64) background block. If this condition is achieved, it is decided that this block is **background**. For non-background block (condition not achieved), it must be decide if this block is natural image or text. So the test of natural image is applied on the block firstly. The decision criterion is based on the use of standard deviation, DCT coefficient absolute sum, gradient vector, and active pixel features. Classification will be achieved by comparing them to the calculated thresholds to make a suitable decision for (64 * 64) image block. If this condition is achieved, it decided that this block is a **natural image** and if not, it must be decided if this block is text or unclassified block. As a final step the test of text is applied on the block. The decision criteria is based on the use of standard deviation, DCT coefficient absolute sum, sum of second derivatives of average intensity (Dx, Dy), and active pixel features. Classification will be achieved by comparing them to the calculated thresholds to make a suitable decision for (64 * 64) text block. If this condition is achieved, it decided that this block is a **text** and if not, it must be decided that this block is **unclassified**.
The steps of the 1st pass segmentation will be as follow:-

- If $\sigma < \sigma_1$ (64) & dct<dct_1 (64) then the block is considered as **background**. Where σ1 (64) & dct_1 (64) are thresholds calculated to make a suitable decision for (64 * 64) background block.
- If dct > dct_2(64) & $\sigma > \sigma_2$(64) & ($\alpha < \alpha_1$(64) or gc < gc_1(64)) then the block is considered as a **natural image**. Where σ_2 (64), dct_2 (64), α_1 (64) & gc_1 (64) are thresholds calculated to make a suitable decision for (64 * 64) image block.
- If dct_1(64) < dct < dct_2 (64) & dx > dx(64) & dy > dy(64) & $\alpha > \alpha_2$(64) then the block is considered as **text**. Where σ_2 (64), α_2 (64)), dx(64) & dy(64) are thresholds calculated to make a suitable decision for (64 * 64) text block.
- If nothing of this previous decision is valid then the block is considered as **unclassified**.

The advantage of first pass segmentation is to make a good decision for text blocks especially with large fonts and the elapsed time taken is so small.
The drawbacks of the first pass segmentation is the size of the block is slightly large so if a part of block have a class and another part of the same block have another class the decision will be taken for one class only from the two classes. Therefore there are a lot of unclassified blocks leading to a large classification error.
To avoid the drawbacks of the first pass segmentation, we will use the Context-based Classification algorithm as explained in the following section.

3.2 First pass segmentation combined with Context-based algorithm

Firstly we divide the image into non-overlapping (32*32) blocks. Every (64*64) unclassified block from the first pass will generate four (32*32) unclassified blocks. Each (32*32) unclassified block will be further processed by the context-based algorithm [2]. Some modifications are proposed in the microclassifier context-based algorithm. This propsed modification is adding DCT coefficient absolute sum feature to the decision conditions. The unclassified (32*32) block is tested by the microclassifier algorithm. The decision criterion is based on using of the mean of the block and DCT coefficient absolute sum features. Classification will be achieved by comparing them to the calculated thresholds to make a suitable decision for (32 * 32) background. If this condition is achieved, it decided that this unclassified block is a **background** and if not, it must be decided if this block text or nature image block. The test of text is applied on the unclassified block. In this step the decision criterion is based on using DCT coefficient absolute sum feature, the unclassified block is adjacent to text block and the two values of bi-level intensity of the block are close to those of the text block. Classification will be achieved by comparing DCT coefficient absolute sum feature to the calculated threshold to make a suitable decision for (32 * 32) text. If this condition is achieved, it decided that this unclassified block is a **text** and if not, it must be decided if this unclassified block is nature image or still unclassified block. Finally the test of nature image is applied on the unclassified block. The decision criterion is based on the use DCT coefficient absolute sum feature and the unclassified block is adjacent to image block and the value of mean of the block are close to those of the image block. Classification will be achieved by comparing DCT coefficient absolute sum feature to the calculated threshold to make a suitable decision for (32 * 32) nature image. If this condition is achieved, it decided that this unclassified block is a **nature image** and if not, it decided that this block is still **unclassified**.

The steps of the 1st pass with the context-based algorithm will be as follow:-

- If the block is unclassified & its mean < thbg & dct < dct_1 (32) then the block is considered as **background**. Where thbg & dct_1 (32) are thresholds calculated to make a suitable decision for (32 * 32) background block.
- If the block is unclassified & adjacent to text block & the block is a bi-level intensity block [4] and its two values are close to those of the text block & dct>dct_2(32) then the block is considered as **text**. Where dct_2 (32) are thresholds calculated to make a suitable decision for (32* 32) text block.
- If the block is unclassified & adjacent to image block & Mean value of the unclassified is close to that of the adjacent image block & dct_1(32)<dct<dct_2(32) then the block is considered as **image**.
- If nothing of these previous decisions is valid then take a decision to this block is still **unclassified**.

The main drawback of the results of this section is that it doesn't make change for the wrong classification which is taken place in the first pass segmentation. To reduce the classification error and to make an efficient classification we will add another segmentation stage.

3.3 second pass segmentation

In this stage, we divide the image into non-overlapping (8*8) blocks. Every (32*32) unclassified block from the previous stage will generate sixteen (8*8) unclassified blocks. Each (8*8) unclassified block will be further processed by the DCT coefficient absolute sum feature only. The first pass segmentation gives an efficient result for classification of the text regions but it is not good in the classification of natural image or background regions, therefore each of the (8*8) text block will be further processed by the modified microclassifier context based algorithm to check if it has a right decision or not. The second pass segmentation gives more accurate results for the blocks which are unclassified or have error classification. Firstly we test the unclassified (8*8) block, by applying the background test will apply firstly, and the decision criterion is based on the use DCT coefficient absolute sum feature and the value of mean of the block. Classification will be achieved by comparing them to the calculated thresholds to make a suitable decision for (8*8) background image. If this condition is achieved, it decided that this unclassified block is a **background** else, it must be decided that this block is nature image or text. So the test of nature text is applied on the unclassified block. The decision criterion is based on the use DCT coefficient absolute sum feature and the unclassified block is adjacent to text block and the two values of bi-level intensity of the block are close to those of the text block. Classification will be achieved by comparing DCT coefficient absolute sum feature to

the calculated threshold to make a suitable decision for (8*8) text. If this condition is achieved, it decided that this unclassified block is a **text** and if not it must be decided that this unclassified block is **natural image**.
Secondly testing any block doesn't have a background decision, the decision criterion is based on the use DCT coefficient absolute sum feature and the value of mean of the block. Classification will be achieved by comparing them to the calculated thresholds to make a suitable decision for (8*8) background image. If this condition is achieved, it decided that this block changed to be **background** and if not no change will happened to the decision of this block.
Finally testing any block doesn't have a natural image decision, The decision criterion is based on the use DCT coefficient absolute sum feature and the unclassified block is adjacent to image block and the value of mean of the block are close to those of the image block. Classification will be achieved by comparing DCT coefficient absolute sum feature to the calculated threshold to make a suitable decision for (8*8) nature image. If this condition is achieved, it decided that this block changed to be **nature image** and if not no change will happened to the decision of this block.
The steps of the second pass will be as follow:-

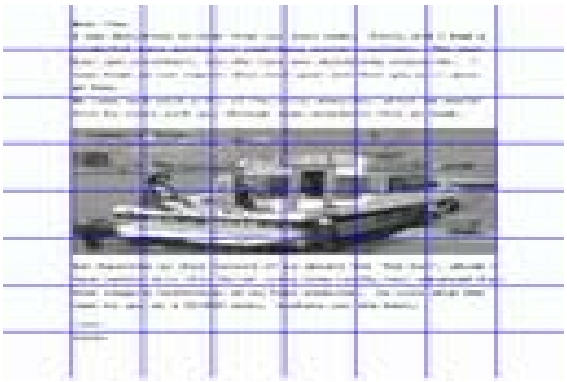
- If the block is unclassified & its mean < thbg & dct < dct_1 (8) then the block is considered as **background**. Where thbg & dct_1 (8) are thresholds calculated to make a suitable decision for (8 * 8) background block.
- If the block is unclassified & adjacent to text block & the block is a bi-level intensity block [4] and its two values are close to those of the text block & dct>dct_2(8) then the block is considered as **text**. Where dct_2 (8) are thresholds calculated to make a suitable decision for (8* 8) text block. If not then the block is considered as **natural image**
- If the block is not background class & dct<dct_1 & its mean < thbg then take a decision to change this block to a **background**.
- If the block is not image & adjacent to image block & Mean value of the unclassified is close to that of the adjacent image block & dct_1(8)<dct<dct_2(8) then take a decision to change this block to a **natural image**.

The drawback of the last pass segmentation is the elapsed time increased but it is still acceptable.
In the next section will show that this proposed scheme will gives more efficient results than the block based algorithms and the DCT based algorithms.

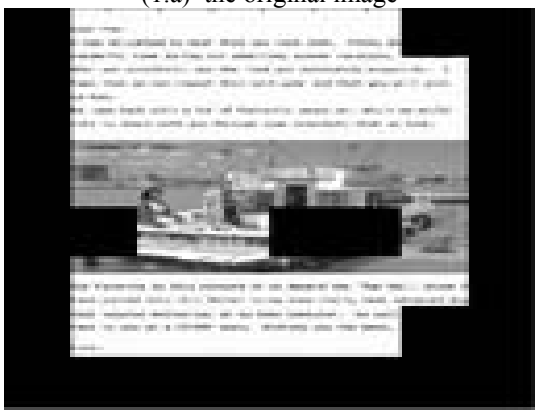
4. Experiments and results

The entire compound image samples are applied on Matlab (7) software where the operating system is Windows XP on a personal computer its processor is three GHZ, the size of samples is in between 512*512 pixels or 1024*1024 pixels. All samples are gray scale, and the extension of all the samples is bitmap. The proposed segmentation algorithm will be called **DCT and block based** algorithm that is because it based on the block based algorithm (six features) and the DCT based algorithm (absolute sum of the DCT coefficient).
In figure (1) a comparison between the first pass segmentation combined with context- based algorithm and the six dimension feature algorithm. The size of compound image sample is 512*512 pixels. In the six feature block based algorithm the text regions is detected well (figure 1.b) but most of the natural image is detected as text (figure 1.d). The classification error is equal to 32%, while the elapsed time taken is 4 seconds. In the first pass segmentation combined with context based algorithm we get the same results for the text regions (figure 1.c) but for the natural image is detected better than six feature block based algorithm(figure 1.e). The classification error is equal to 9% and the elapsed time taken is 4.6 second. From the previous results we can notice that by adding the absolute sum of the DCT coefficient feature to the six dimension features, it gives more efficient results especially for detection the image regions. The modified context-based algorithm gives an accurate results more than before because it doesn't depend only on the adjacent blocks, so it doesn't depend on wrong classification from the first pass segmentation. Therefore the classification error is reduced while the elapsed time is still small.
In figure (2) a comparison between the block based algorithms, DCT based algorithm and the proposed algorithm. The size of compound image sample is 512*512 pixels. In the six feature block based algorithm the text regions is detected well (figure 2.b) but most of the natural image is detected as text (figure 2.e). The classification error is equal to 12%. In the Delta probabilities DCT based algorithm the image and the background are detected well (figure 2.f) but most of the text regions are detected as natural image (figure 2.c). The classification error is equal to 9%. The elapsed time taken for the six feature algorithm is 10 second while the elapsed time taken for the delta probabilities DCT based algorithm is 14 second. In the proposed algorithm the text regions is detected well (figure 2.d) and most of the natural image is also detected well (figure 2.g) the

classification error is equal to 1.8% while the elapsed time taken is 16 second. From the previous results we can notice that the proposed algorithm has more efficient results and the classification error is reduced and becomes more acceptable than the six dimension feature algorithm and the DCT based algorithms. The reason behind is the using of the six dimension features which is mentioned above with the DCT absolute some coefficient feature in the first pass segmentation and adding the modified context-based algorithm. Also by comparing the final results of the proposed algorithm with any of the block based algorithms it is shown that the classification error of the proposed algorithm is less than the block based algorithms that is because the using of DCT absolute some coefficient feature which is improved the results and support the six features to give more efficient segmentation and classification results. Since it was proved that the six dimension feature algorithm gives more efficient results than wavelet based algorithm in [1], therefore we can say the proposed segmentation algorithm is better than the wavelet based algorithm

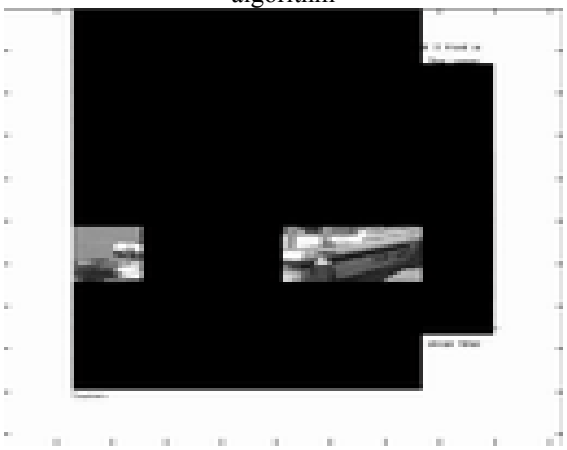


(1.a) the original image

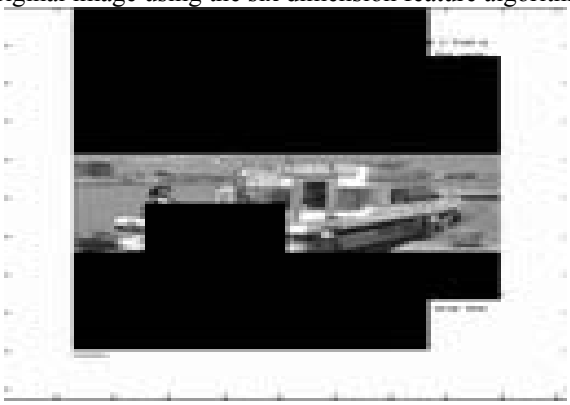


(1.b) text segmented from the original image using the six dimension feature algorithm.

(1.c) text segmented from the original image using the first pass algorithm combined with context- based algorithm

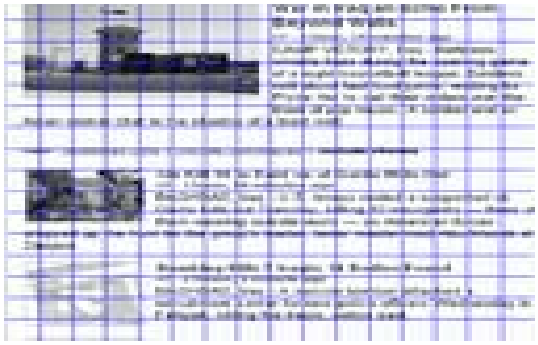


(1.d) image and background segmented from the original image using the six dimension feature algorithm

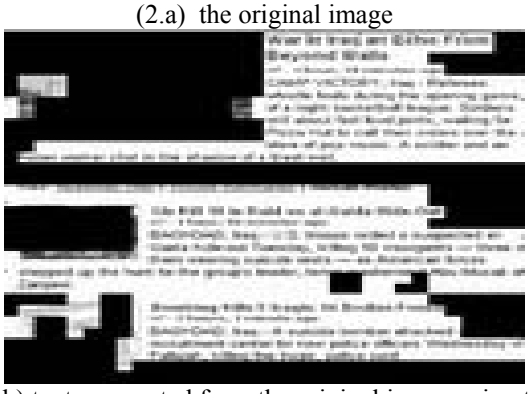


(1.e) image and background segmented from the original image using the first pass algorithm combined with context- based algorithm

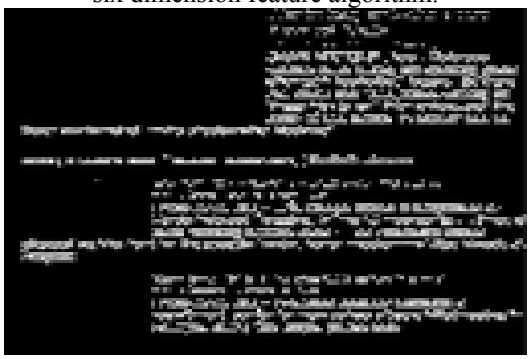
Figure1. Comparison between the first pass algorithm combined with context- based algorithm and the six dimension feature algorithm



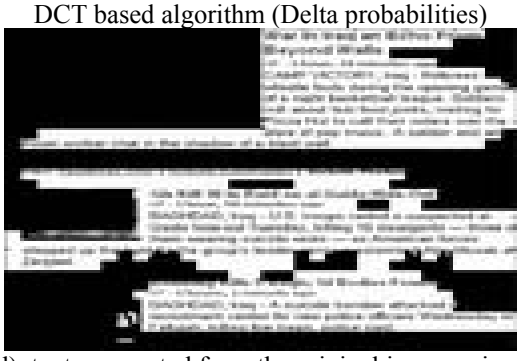
(2.a) the original image



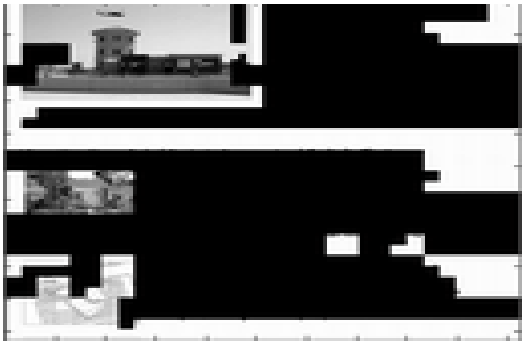
(2.b) text segmented from the original image using the six dimension feature algorithm.



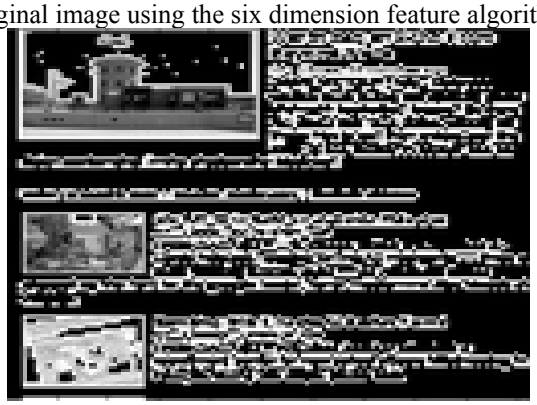
(2.c) text segmented from the original image using the DCT based algorithm (Delta probabilities)



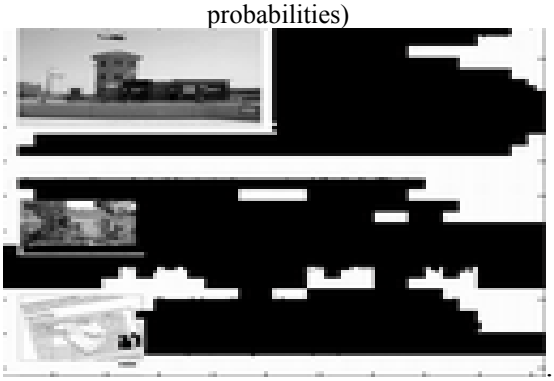
(2.d) text segmented from the original image using the proposal algorithm



(2.e) image and background segmented from the original image using the six dimension feature algorithm



(2.f) image and background segmented from the original image using the DCT based algorithm (Delta probabilities)



(2.g) image and background segmented from the original image using the proposal algorithm

Figure2. Comparison between the block based algorithm, DCT based algorithm and the proposal algorithm

5. Conclusions

In this paper we present a new scheme to make an efficient segmentation for the compound image, this scheme is called DCT and block based algorithm. It uses

the advantage of six features works together. This scheme depend on the DCT based algorithm by using the absolute sum of the DCT coefficients. The scheme using also the context – based algorithm which helps to give more performance results. The proposed algorithm starts by partitioning the image to 64*64 blocks which help to segment the text regions especially which have large fonts, and end by partition the image to 8*8 blocks which is good in segmentation of natural images especially by using DCT absolute sum coefficients.
The classification error percentage of this scheme is less than 3%. A comparison have been happened between this scheme and the other schemes such as block based segmentation algorithms (six dimension features algorithm) and DCT based algorithms (Delta probabilities). The results proved that the proposed scheme (DCT-block scheme) is gives more efficient results and less classification error with slightly increasing in the elapsed time taken.

6. References

[1] Kush R. Varshney, "Block-segmentation and Classification of Grayscale Postal Images" *School of Electrical and Computer Engineering Cornell University, Ithaca, NY 1485,* may 2003

[2] J. Li and R.M. Gray, "Context-based multiscale classification of document images using wavelet coefficient distributions," *IEEE Trans. Image Processing*, vol. 9, pp. 1604-1616, Sept. 2000.

[3] Mark Kalman, Isaac Keslassy_, Daniel Wang, and Bernd Girod, "Classifications of compound images based on transform coefficient likelihood" *Information Systems Laboratory, Department of Electrical Engineering Stanford University, Stanford, CA 94305,* july2002

[4] A.Said and A. Drukarev, "Simplified segmentation for compound image compression," *Proc. 1999 Int. Conf. on Image Processing*, vol. 1, pp. 229-233.

[5] A.Said, "Compression of compound images and video for enabling rich media in embedded systems," *SPIE Visual Commun. Image Processing Conf.,* SPIE Proc. vol. 5308, pp. 69–82, Jan. 2004.

[6] Amir Said "Efficient and reliable dynamic quality control for compression of compound document" *Proc. ICIP*, 2004.

[7] Tony Lin and Pengwei H, Sang Uk Lee "Efficient Coding of Computer Generated Compound Images"proc.ICIP, 2005

[8] J.D. Foley, A. van Dam, S.K. Feiner, and J.F. Hughes, *Computer Graphics, Principles and Practices*, Addison-
Wesley Pub. Co., Reading, MA, 1990.

[9] T.H. Cormen, C.E. Leiserson, and R.L. Rivest, *Introduction to Algorithms,* MIT Press, Cambridge, MA, 1990.

[10] R.C. Gonzalez and R.E. Woods, *Digital Image Processing*, Addison-Wesley Pub. Co., Reading, MA, 1992.

[11] J.M. Gilbert and R.W. Brodersen, "A lossless 2-D image compression technique for synthetic discrete-tone images," *Proc. Data Compression Conf.,* pp. 359–368, March 1998

[12] H. Cheng and C.A. Bouman, "Multilayer document compression algorithm," *Proc. IEEE Int. Conf. Image Processing,* vol. 1, pp. 244–248, Oct. 1999.

[13] R.L. de Queiroz, "Compression of Compound Documents," *Proc. IEEE Int. Conf. Image Processing,* vol. 1,
pp. 24–28, Oct. 1999.

[14] P. Haffner, Y. LeCun, L. Bottou, P. Howard, P. Vincent, and B. Riemers, "Color documents on the Web with
DjVu," *Proc. IEEE Int. Conf. Image Processing,* Oct. 1999.

[15] D.A. Tompkins and F. Kossentini, "A fast segmentation algorithm for bi-level image compression using JBIG2,"
Proc. IEEE Int. Conf. Image Processing, Oct. 1999.

[16] R.L. de Queiroz, Z. Fan, and T.D. Tran, " Optimizing block-threshold segmentation for MRC compression," *Proc.*
IEEE Int. Conf. Image Processing, vol. 2, pp. 597–600, Sept. 2000.