DWT Based Text Localization

Chung-Wei Liang and Po-Yueh Chen*

*Department of Computer Science and Information Engineering,
Chaoyang University of Technology,
Wufeng, Taichung county 413, Taiwan, R.O.C.*

Abstract: This paper presents an efficient yet simple method to extract text regions from static images or video sequences. The operation speed of Haar discrete wavelet transform (DWT) operates the fastest among all wavelets because its coefficients are either 1 or -1. This is one of the reasons we employ Haar DWT to detect edges of candidate text regions. The resulted detail component sub-bands contain both text edges and non-text edges. However, the intensity of the text edges is different from that of the non-text edges. Therefore, we can apply thresholding to preliminary remove the non-text edges. Text regions are composed of vertical edges, horizontal edges and diagonal edges. Morphological dilation operators are applied to connect isolated text edges of each detail component sub-band in a transformed binary image. According to the experiment results, real text regions are the overlapped portion of three kinds of dilated edges. Hence, we can apply the logical AND operator to three kinds of dilated edges and obtain the final text regions correctly.

Keywords: text extraction; Haar DWT; Thresholding; Morphological operator; Logical AND operator.

1. Introduction

Texts in images and video sequences provide highly condensed information about the contents of the images or videos sequences and can be used for video browsing/retrieval in a large video database. Although texts provide important information about images or video sequences, it is not an easy problem to detect and segment them. Texts extraction is not easy for the following reasons. First of all, text sizes may change from small to big and text fonts may vary in a wide range as well. Secondly, texts present in an image or a video sequence may have multiple colors and appear in a very cluttered background. Many papers about the extraction of texts from static image or video sequence have been published in recent years. Those methods for texts extraction can be classified as either component-based or texture-based. Using component-based texts extraction methods, text regions are detected by analyzing the edges of the candidate regions or homogenous color/grayscale components that contain the characters. For example, Park et al. [1] detected eight orientations of edge pixels in the documents using prewitt masks. Edge pixels can be classified as either axial directional or diagonal directional. They dilate two kinds of edge pixels using morphological dilation operators. Then logical AND is applied to these two edges to obtain the real text regions. Zhong et al. [2] located bounding boxes

* Corresponding author: e-mail: pychen@mail.cyut.edu.tw

Accepted for Publication: Feb. 20, 2004

around text components using the horizontal spatial variance. Candidate text regions have higher horizontal spatial variance than that of non-text regions. In each candidate text region, connected components are of the same color. They determined the color of texts and locate text components in each candidate text region. Finally, real text components are filled in each candidate text region. Chen et al. [3] detected vertical edges and horizontal edges in an image and dilated two kinds of edges using different dilation operators. The logical AND operator is performed on dilated vertical edges and dilated horizontal edges to obtain candidate text regions. Real Text regions are then identified using support vector machine. Text regions usually have a special texture because they consist of identical character components. These components also contrast the background and hence text regions have a periodic horizontal intensity variation due to the horizontal alignment of characters. As a result, text regions can be segmented using texture features. For example, Stephen et al. [4] did the segmentation and labeling of block using the connected component analysis. Paul et al [5] segmented and classified texts in a newspaper by generic texture analysis. Small masks are applied to obtain local textural characteristics.

Most of the text extraction methods were applied to uncompressed images. Few of them proposed to extract texts in the compressed version of images. Zhong et al. [6] extracted captions from the compressed videos (MPEG video and JPEG image) based on Discrete Cosine Transform (DCT). DCT detects edges in different directions from the candidate image. Edge regions containing texts are then detected using a threshold afterward. Acharyya et al. [7] segmented texts in the document images based on wavelet scale-space features. The method used the M-band wavelet which decomposes an image into some M×M bandpass channels so as to detect the text regions easily. The intensity of the candidate text edges are used to recognize the real text regions in an M-band image.

In this paper, we proposed an efficient method that extracts text regions in video sequences or images by using Haar discrete wavelet transform (Haar DWT), thresholding and morphological operators. First of all, Haar DWT detects three kinds of edges and preliminarily removes the non-text regions in the detail component sub-bands with the aid of the thresholing technique. Then text edges and non-text edges are distinguished successfully by morphological operators and the logical AND operator. The proposed extraction algorithm is described with details in Section 2. In Section 3, experiment results are displayed for some complicated images and videos. We choose samples with both text regions and graphical regions so as to demonstrate the efficiency of the proposed method. Finally, we conclude in Section 4.

2. DWT based extraction

In this section, we present a method to extract texts in images or video sequences using Haar discrete wavelet transform (Haar DWT). The edges detection is accomplished by using 2-D Haar DWT and some of the non-text edges are removed using thresholding. Afterward, we use different morphological dilation operators to connect the isolated candidate text edges in each detail component sub-band of the binary image.

Although the color component may differ in a text region, the information about colors does not help extracting texts from images. If the input image is a gray-level image, the image is processed directly starting at discrete wavelet transform. If the input image is colored, its RGB components are combined to give an intensity image Y as follows:

$$Y = 0.299R + 0.587G + 0.114B \tag{1}$$

Image Y is then processed with discrete wavelet transform and the whole extraction algorithm afterward. If the input image itself is stored in the DWT compressed form, DWT operation can be omitted in the proposed algorithm. The flow chart of the proposed algorithm is shown in Figure 1.

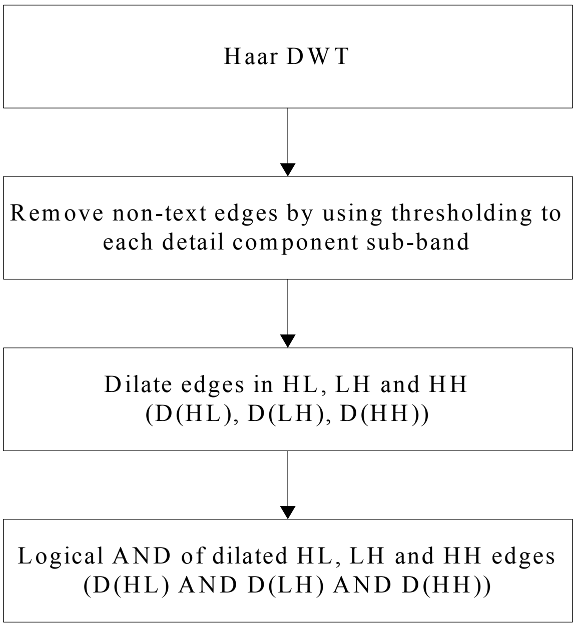

Haar DWT
Remove non-text edges by using thresholding to each detail component sub-band
Dilate edges in HL, LH and HH (D(HL), D(LH), D(HH))
Logical AND of dilated HL, LH and HH edges (D(HL) AND D(LH) AND D(HH))

Figure 1. Flow chart of the proposed text extraction algorithm

2.1. Haar discrete wavelet transform

The discrete wavelet transform is a very useful tool for signal analysis and image processing, especially in multi-resolution representation [8]. It can decompose signal into different components in the frequency domain. One-dimensional discrete wavelet transform (1-D DWT) decomposes an input sequence into two components (the average component and the detail component) by calculations with a low-pass filter and a high-pass filter [9]. Two-dimensional discrete wavelet transform (2-D DWT) decomposes an input image into four sub-bands, one average component (LL) and three detail components (LH, HL, HH) as shown in Figure 2.

In image processing, the multi-resolution of 2-D DWT has been employed to detect edges of an original image. The traditional edge detection filters can provide the similar result as well. However, 2-D DWT can detect three kinds of edges at a time while traditional edge detection filters cannot. As shown in Figure 3, the traditional edge detection filters detect three kinds of edges by using four kinds of mask operators. Therefore, processing times of the traditional edge detection filters is slower than 2-D DWT.

LL	HL
LH	HH

Figure 2. The result of 2-D DWT decomposition

Figure 4 (a) shows a gray level image. The 9-7 taps DWT filters decompose this gray image into four sub-bands as shown in Figure 4 (b). As we can see, three kinds of edges present in the detail component sub-bands but look unobvious (very small coefficients). If we replace the 9-7 taps DWT filters with Haar DWT, the detected edges become more obvious and the processing time decreases.

The operation for Haar DWT is simpler than that of any other wavelets. It has been applied to image processing especially in multi-resolution representation [10]. Harr DWT has the following important features [11].

1. Haar wavelets are real, orthogonal, and symmetric.
2. Its boundary conditions are the simplest among all wavelet-based methods.
3. The minimum support property allows arbitrary spatial grid intervals.
4. It can be used to analyze texture and detect edges of characters.
5. The high-pass filter and the low-pass filter coefficient is simple (either 1 or –1).

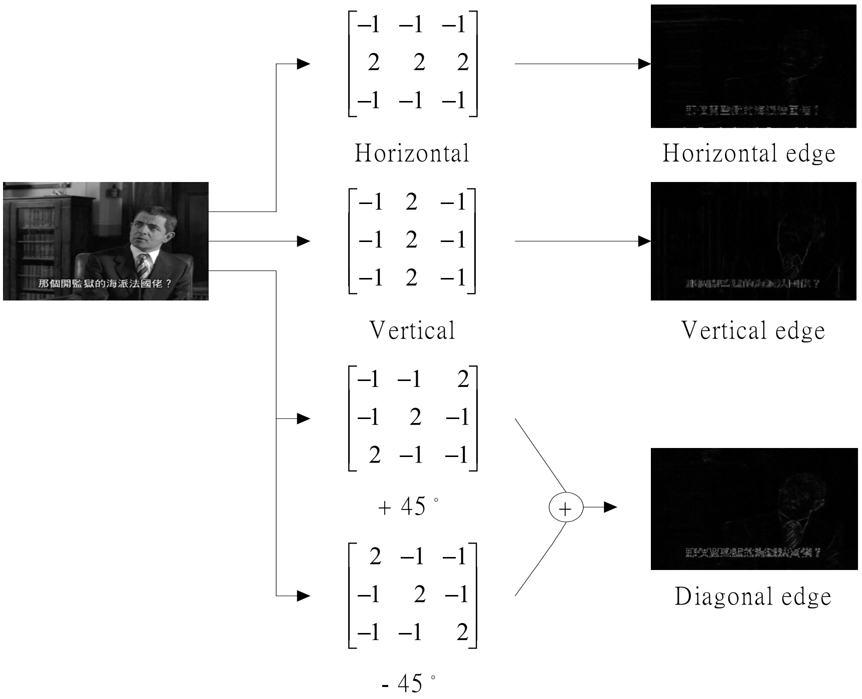
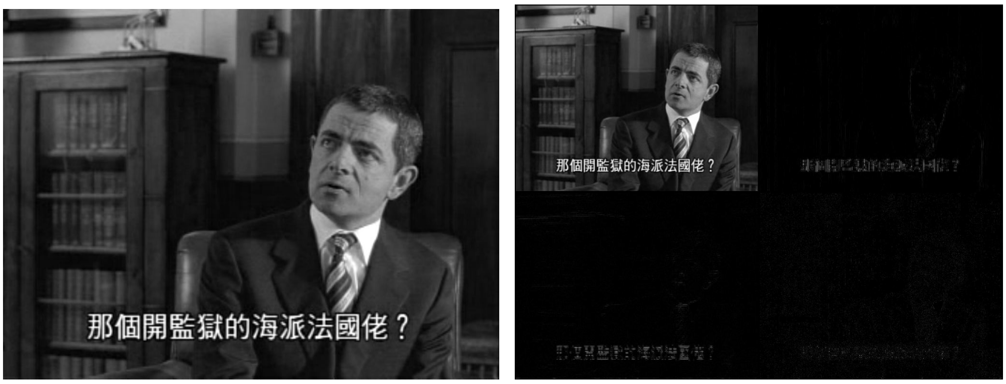


Figure 3. Traditional edge detection using mask operation



(a)

(b)

Figure 4. (a) Original gray image (b) DWT coefficients

Figure 5 (a) shows the example of a 4×4 gray-level image. The wavelet coefficients can be obtained in gray-level image using addition and subtraction. 2-D DWT is achieved by two ordered 1-D DWT operations (row and column). First of all, we perform the row operation to obtain the result shown in Figure 5 (b). Then it is transformed by the column operation and the final resulted 2-D Haar DWT is shown in Figure 5 (c). 2-D Haar DWT decomposes a gray-level image into one average component sub-band and three detail component sub-bands.

$$\begin{bmatrix} A & B & C & D \\ E & F & G & H \\ I & J & K & L \\ M & N & O & P \end{bmatrix} \qquad \begin{bmatrix} (A+B) & (C+D) & (A-B) & (C-D) \\ (E+F) & (G+H) & (E-F) & (G-H) \\ (I+J) & (K+L) & (I-J) & (K-L) \\ (M+N) & (O+P) & (M-N) & (O-P) \end{bmatrix}$$

(a) (b)

$$\begin{bmatrix} (A+B)+(E+F) & (C+D)+(G+H) & (A-B)+(E-F) & (C-D)+(G-H) \\ (I+J)+(M+N) & (K+L)+(O+P) & (I-J)+(M-N) & (K-L)+(O-P) \\ (A+B)-(E+F) & (C+D)-(G+H) & (A-B)-(E-F) & (C-D)-(G-H) \\ (I+J)-(M+N) & (K+L)-(O+P) & (I-J)-(M-N) & (K-L)-(O-P) \end{bmatrix}$$

(c)

Figure 5. (a) The original image (b) the row operation of 2-D Haar DWT (c) the column operation of 2-D Haar DWT

In those three detail components of a Haar DWT image, we can obtain various features about the original image as follows:

1. Average components are detected by the LL sub-band;
2. Vertical edges are detected by the HL sub-band;
3. Horizontal edges are detected by the LH sub-band;
4. Diagonal edges are detected by the HH sub-band.

For example, the gray-level image shown in Figure 4 (a) is decomposed into 2-D Haar DWT as shown in Figure 6. We can detect candidate text edges in the original image from those three detail component sub-bands (HL, LH and HH) in Figure 6.

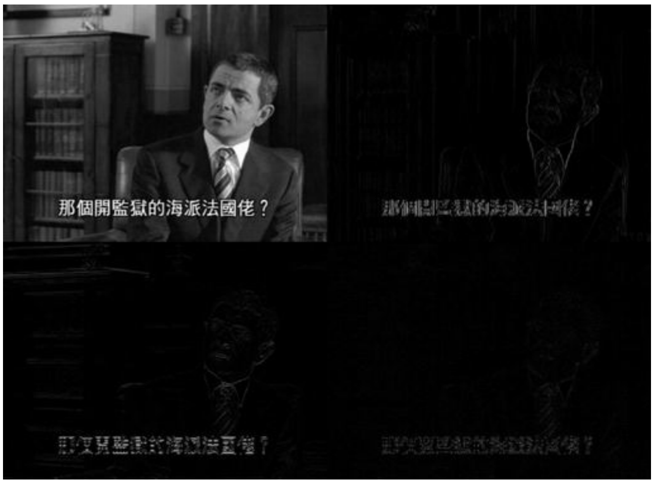


Figure 6. 2-D Haar discrete wavelet transform image

Chen and Liao [12] presented the segment-matrix algorithm for Haar DWT to decrease the processing time of DWT operations. The method produces the same results as traditional Haar DWT with a much faster speed. Hence, we apply the segment-matrix algorithm to decompose an original gray-level image into four sub-bands. Figure 7 (a) shows an example of a 4×4 gray-level image. It is segmented into 4 2×2 sub-blocks as shown in Figure 7 (b). Then each 2×2 sub-block is performed with the z-scan operation and we obtain 4 1×4 sub-blocks as shown in Figure 7 (c). The Haar DWT filter coefficient matrix (presented in Figure 7 (d)) is multiplied by the matrix shown in Figure 7 (c) and then the result of 2-D DWT is obtained in Figure 7 (e).

After the Haar DWT, the detected edges include mostly text edges and some non-text edges are presented in the 3 detail component sub-bands. In next subsection, we employ dynamic thresholding to preliminarily remove those non-text edges in the detail component sub-bands.

$$\begin{bmatrix} A & B & C & D \\ E & F & G & H \\ I & J & K & L \\ M & N & O & P \end{bmatrix}$$

(a)

$$\begin{bmatrix} \begin{bmatrix} A & B \\ E & F \end{bmatrix} & \begin{bmatrix} C & D \\ G & H \end{bmatrix} \\ \begin{bmatrix} I & J \\ M & N \end{bmatrix} & \begin{bmatrix} K & L \\ O & P \end{bmatrix} \end{bmatrix}$$

(b)

$$\begin{bmatrix} [A & B & E & F] \\ [C & D & G & H] \\ [I & J & M & N] \\ [K & L & O & P] \end{bmatrix}$$

(c)

$$\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix}$$

(d)

$$\begin{bmatrix} A+B+E+F & C+D+G+H & A-B+E-F & C-D+G-H \\ I+J+M+N & K+L+O+P & I-J+M-N & K-L+O-P \\ A+B-E+F & C+D-G+H & A-B-E-F & C-D-G-H \\ I+J-M+N & K+L-O+P & I-J-M-N & K-L-O-P \end{bmatrix}$$

(e)

Figure 7. (a) the original image (b) the original image to change into 2×2 sub-blocks (c) the z-scan result of the 2×2 sub-blocks (d) the filter coefficient matrix (e) the correct result of 2-D Haar DWT

2.2. Thresholding

Thresholding is a simple technique for image segmentation. It distinguishes the image regions as objects or the background. Although the detected edges are consist of text edges and non-text edges in every detail component sub-band, we can distinguish them due to the fact that the intensity of the text edges is higher than that of the non-text edges. Thus, we can select an appropriate threshold value and preliminarily remove the non-text edges in the detail component sub-bands. In this subsection, we employ dynamic thresholding [13] to calculate the target threshold value *T*. The target threshold value is obtained by performing an equation on each pixel with its neighboring pixels. We employ two mask operators to obtain such an equation and then calculate the threshold value for each pixel in the 3 detail sub-bands. Basically, the dynamic thresholding method obtains different target threshold values for different images.

Each detail component sub-band ***es*** is then compared with *T* to obtain a binary image (***e***).

The threshold T is determined by

$$T = \frac{\sum (es(i,j) \times s(i,j))}{\sum s(i,j)} \qquad (2)$$

where

$$s(i, j) = Max(\left|g1 ** es(i, j)\right|, \left|g2 ** es(i, j)\right|) \qquad (3)$$

and

$$g1 = [-1 \ \ 0 \ \ 1], g2 = [-1 \ \ 0 \ \ 1]^t \qquad (4)$$

In Eq. (3), "**" denote two-dimensional liner convolution.

Figure 8 shows the example of a 5×5 detail component sub-band *(es)*. We calculate S(P8) as an example to demonstrate the definition of Eqs. (3) and (4).

$$S(P8) = max\left(\left|P9 - P7\right|, \left|P13 - P3\right|\right) \qquad (5)$$

Applying similar operations to each pixel, we obtain all the S(i, j) for each detail component sub-band. After that, we can apply Eq. (2) to compute T and the binary edge image (***e***) is then given by

$$e(i,j) = \begin{cases} 255, & if \ es(i,j) > T \\ 0 \ \ , & otherwise \end{cases} \qquad (6)$$

The resulted binary image, as shown in Figure 9, is mostly consisted of text edges and very few non-text edges.

$$\begin{bmatrix} P1 & P2 & P3 & P4 & P5 \\ P6 & P7 & P8 & P9 & P10 \\ P11 & P12 & P13 & P14 & P15 \\ P16 & P17 & P18 & P19 & P20 \\ P21 & P22 & P23 & P24 & P25 \end{bmatrix}$$

Figure 8. 5×5 detail component sub-band *(es)*

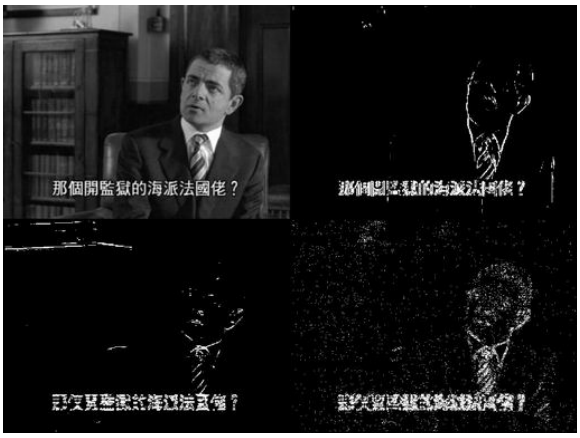


Figure 9. Binary image of detail component sub-band

2.3. Text region extraction

In this subsection, we use morphological operators and the logical AND operator to further remove the non-text regions. In text regions, vertical edges, horizontal edges and diagonal edges are mingled together while they are distributed separately in non-text regions.

Since text regions are composed of vertical edges, horizontal edges and diagonal edges, we can determine the text regions to be the regions where those three kinds of edges are intermixed. Text edges are generally short and connected with each other in different orientation. In Figure 10, we use different morphological dilation operators to connect isolated candidate text edges in each detail component sub-band of the binary image.

In this paper, 3×5 for horizontal operators, 3×3 for diagonal operators and 7×3 for vertical operators as in shown Figure 11 are applied. The dilation operators for the three detail sub-bands are designed differently so as to fit the text characteristics. The logical AND is then carried on three kinds (vertical, horizontal and diagonal) of edges after morphological dilation. This process is indicated in Figure 12. Since three kinds of edge regions are intermixed in the text regions, overlapping appears a lot after the morphological dilation due to the expansion of each single edge. On the contrary, only one kind of edge region or two kinds of edge regions exist separately in the non-text regions and hence there is no overlapping even after the dilation. Therefore, the AND operator helps us to obtain the candidate text regions as shown in Figure 13 (a). Sometimes the text candidate regions may contain some non-text component regions which are too large or too small. By limiting the block size, we obtain the final text regions. Each text region has a moderate size w × h (pixels) in a candidate text region image.

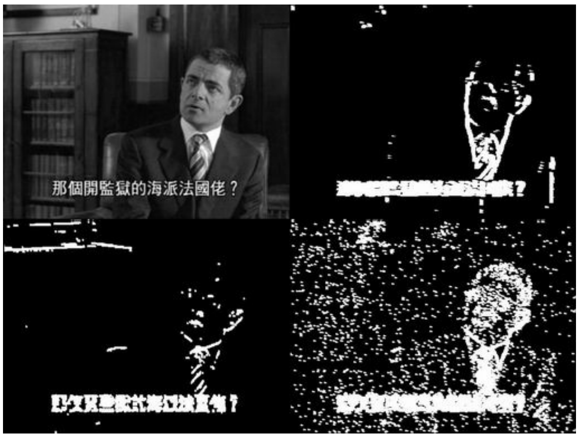


Figure 10. The dilated image of three binary regions

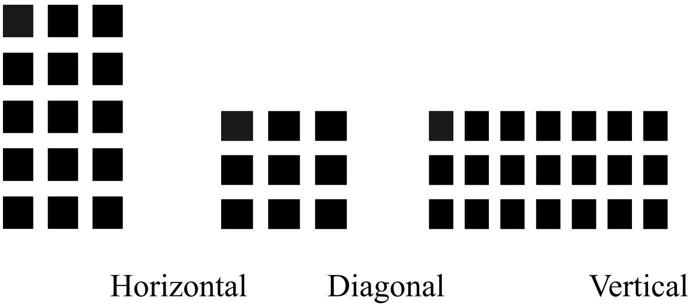


Figure 11. Horizontal, Diagonal and Vertical edges dilation operators

The minimum text block size is determined as follows:

$$width > 100(pixels), height > 35(pixel) \qquad (7)$$

Removing the candidate text regions smaller than this limit, the final text region is shown in Figure 13 (b).

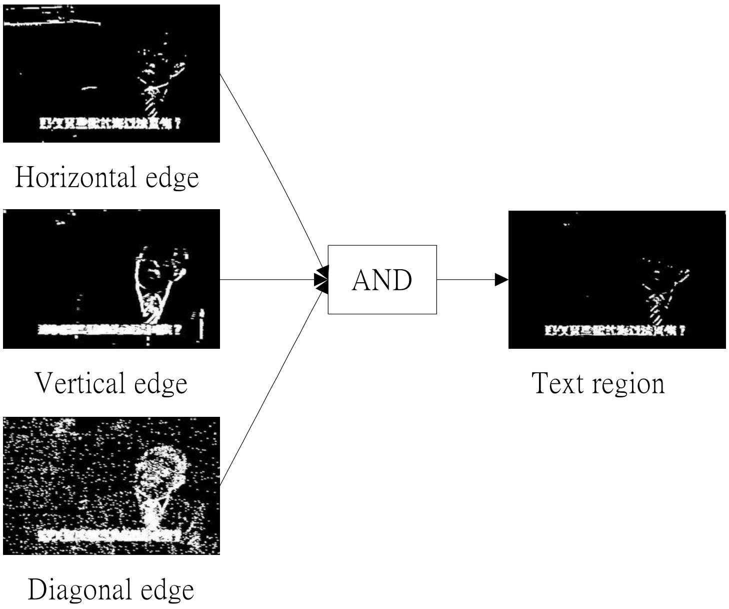
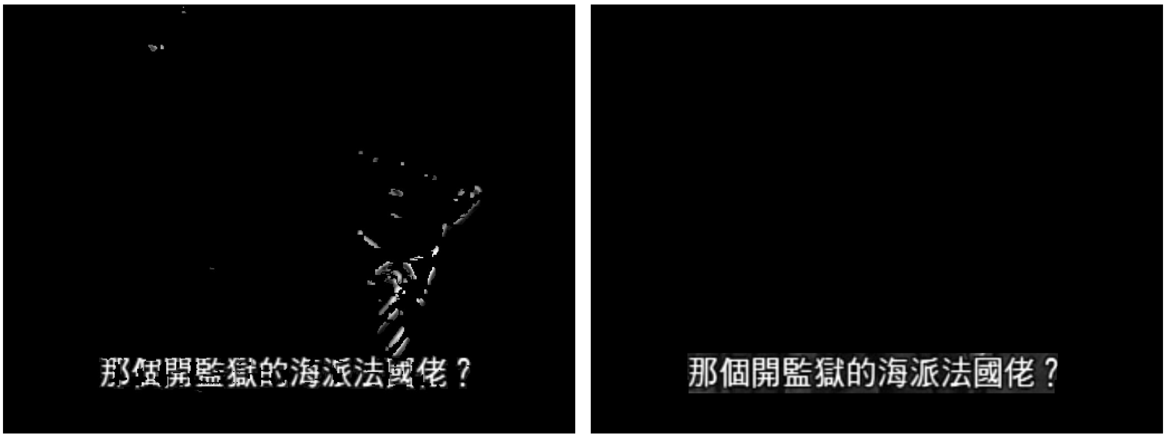


Figure 12. Text extraction by using the logical AND operator



(a) (b)

Figure 13. (a) The candidate text region (b) the extracted real text region

3. Experiment results

Experiments were carried out on video of movies and images. Each video frame or image has 1024×768 resolution in BMP or MPEG format. Since they are colored, we convert them into gray-level before applying text localization. In Figure 14, several experiments were preformed to locate the text regions in video frames. The proposed method can correct locate text regions in complex images. The processing time of the proposed is faster than other methods as shown in Table 1. When the semented-matrix Haar DWT is applied, the speed performance is improved by a portion more than 50%. In Table 2, we see the correct detecting rate of the proposed method and that of the other

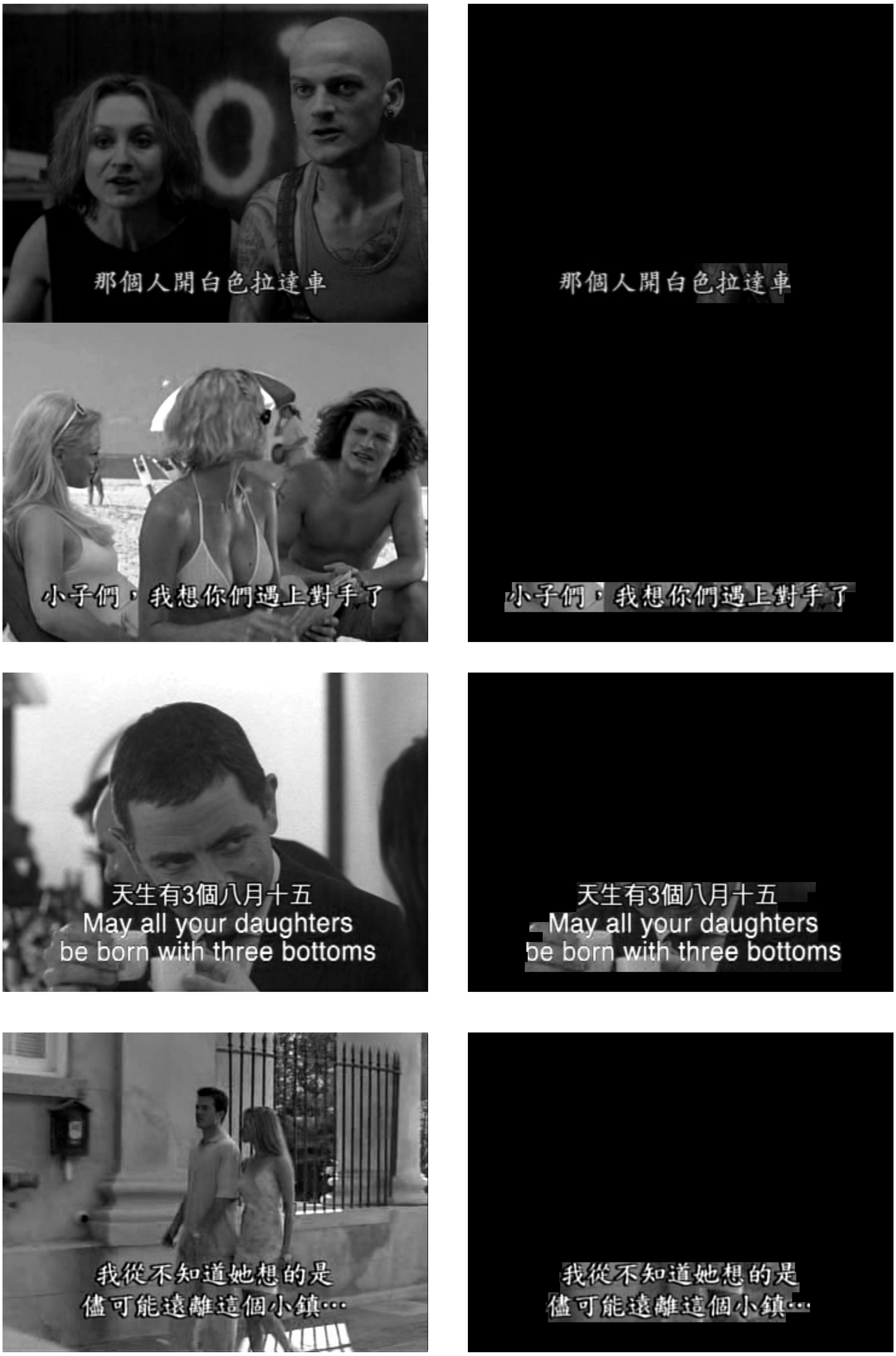


Figure 14. Examples of text localization

methods are almost the same. The correct detecting rate is defined as the ratio of number of pixels in the real text regions to the number of pixels in the detected regions (Since text components which are too small or too large are considered as noise, basically all methods successfully detect 100% of the text regions). Based on the experiment results, we proof the proposed method is effective and efficient for detecting the captions in video frames.

Table 1. Processing time of text localization methods

	Processing time of text localization
Traditional edge detection	0.906(s)
9-7 taps DWT	0.875(s)
Traditional Haar DWT	0.781(s)
Segment-matrix Haar DWT	0.39(s)

Table 2. Average Correct rate and average Error rate

	Average correct rate	Average error rate
Traditional edge detection	98.7604%	1.2396%
9-7 taps DWT	98.7843%	1.2157%
Traditional Haar DWT	98. 5719%	1.4281%
Segment-matrix Haar DWT	98.5719%	1.4281%

4. Conclusion

A direct and efficient text extraction scheme is proposed using Haar DWT, the morphological dilation operators and the logical AND operator. We integrate these mathematical tools to detect the text regions from complicated images. In order to fit the text characteristics, the dilation operators for the three detail sub-bands are designed differently. According to the experiment results, the proposed scheme is proved to be efficient for extracting text regions from the images or video sequences.

References

[1] Park, C. J., Moon, K. A., Oh, Weon-Geun, and Choi, H. M. 2000. An efficient of character string positions using morphological operator. *IEEE International Coanference on Systems, Man, and Cybernetics,* 3, 8-11: 1616-1620.

[2] Zhong, Yu., Karu, K., and Jain, A.K. 1995. Locating text in complex color images. *Proceedings of the Third International Conference on Document Analysis and Recognition,* 1, 14-16: 146-149.

[3] Chen, Datong, Bourlard, H., and Thiran J. P., 2001. Text identification in complex background using SVM. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Proceedings of the 2001*, 2, 8-14: 621-626.

[4] Lam, S. W., Wang, D., and Srihari, S. N., 1990. Reading newspaper text. *International Conference on Pattern Recognition Proceedings, 10th.* I, 16-21:

703-705.

[5] Williams, P. S. and Alder, M. D. 1996. Generic texture analysis applied to newspaper segmentation. *IEEE International Conference on Neural Networks,* 3, 3-6: 1664-1669.

[6] Zhong, Yu., Zhang, Hongjiang., and Jain, A. K. 2000. Automatic caption localization in compressed video. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 22, 4: 385 –392.

[7] Acharyya, M. and Kundu, M. K. 2002. Document image segmentation using wavelet scale-space features. *IEEE Transactions on Circuits and Systems for Video Technology*, 12, 12: 1117 –1127.

[8] Mallat, S. G. 1989. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11, 7: 674-693.

[9] Acharya, Tinku., Chen, Po-Yueh. 1998. VLSI implementation of a DWT architecture. *Proceedings of the IEEE International Symposium on Circuits and Systems*, 2: 272-275.

[10] Grochening, K. and Madych, W. R. 1992. Multiresoultion analysis, Haar bases, and self-similar tilings of R^n . *IEEE Transactions on Information Theory,* 38, 2.

[11] Fujii, Masafumi., Wolfgang, J. R., and Hoefer. 2001. Filed-Singularity correction in 2-D time-domain Haar-wavelet modeling of waveguide components. *IEEE Transactions on Microwave Theory and Techniques*, 49, 4.

[12] Chen, P. Y. and Liao, E. C. 2002. A new algorithm for Haar discrete wavelet transform. *IEEE International Symposium on Intelligent Signal Processing and Communication Systems*, 21, 24: 453-457.

[13] Hasan, M. Y. and Lina J, Y. K. Morphological text extraction from image. *IEEE Transactions on Image Processing*, 9, 11: 1978 -1983.