Text Localization of the JPEG images and MPEG Compressed videos

Nikolay N. Neshov[1], Ivo R. Draganov[2], Darko Brodic[3]
[1,2] Faculty of Telecommunications
8 Kliment Ohridski Blvd., 1000 Sofia
Bulgaria
nneshov@tu-sofia.bg, idraganov@tu-sofia.bg

[3]University of Belgrade, Technical Faculty in Bor
V.J. 12, 19210 Bor, Serbia
dbrodic@tf.bor.ac.rs

ABSTRACT: *In the JPEG compressed images and I-frames of the MPEG compressed videos, the text region extraction is challenging which can be carried out through localization. We in this paper have identified the text location with the text regions using DCT compression which is in contrast to conservative full decompression of the videos. Due to the advantage of less decoding, the compression process seems to be more faster. This process ensures text detection in large sized images. While we applied the algorithm we are able to arrive at effective results.*

Keywords: Text Extraction, DCT, JPEG, MPEG

Received: 24 February 2022, Revised 15 April 2022, Accepted 14 May 2022

DOI: 10.6025/jmpt/2022/13/3/67-73

1. Introduction

Many digital images are recorded, transferred, and processed in a compressed format. Thus, a faster text information detection system can be achieved if text extraction can be done without decompression. Extraction of this information involves detection, localization, extraction, enhancement, and recognition of the text from a given image. However, variations of text due to differences in size, style, orientation, and alignment, as well as low image contrast and complex background make the problem of automatic text extraction extremely challenging.

Some features, particularly those based on the image's spatial frequency, have been proved to be very useful. In [1] is developed algorithm for detecting, binarizing, and tracking caption text ingeneral-purpose MPEG video. An advantage of the frequency-based methods is that they can realize a fast text detection in the DCT compressed domain or Wavelet Transform (WT) [2, 3]. Many authors use the DCT coefficients of lower frequency, while others use the Wavelet coefficients of higher frequency.

Usually, the evaluations of the features for text detection have been made mainly through the evaluations of the final results of text region detection [4, 5, 6]. Analyses of the features for text/background separation have not been studied enough. No

research has yet been made for finding a frequency band that is potentially good for text region detection.

Some works suggest the weighted DCT coefficient based text detection, mainly empirically defined [7]. Fisher's discriminant criterion [8] has been used for optimizing the text features as well.

We focus on analyses and evaluations of the DCT-based features and post-filtration of the non-text DCT blocks. In Section 2, is introduced the proposed text detection method using DCT-based features in low frequency band, a filtration method in 6 steps for candidate-text regions and the proposed algorithm's block diagram are presented. We analyze and include experimental results is Sections 3 and in section 4 conclusion is done.

2. Algorithm Description

All operations are performed in the Discrete Cosine Transform (DCT) domain for high speed processing. The basic algorithm is accomplished as follows: For each DCT block 8x8 a sum of specially chosen AC coefficients S_{AC} (shown in Figure 1) are used to compute a specific feature called text energy of the particular block [1]:

$$S_{AC}=\sum_{i=1}^{5}AC_i+\sum_{i=8}^{12}AC_i+\sum_{i=16}^{19}AC_i+\sum_{i=24}^{26}AC_i+AC_{32}+AC_{40}. \quad (1)$$

DC	AC_1	AC_2	AC_3	AC_4	AC_5	AC_6	AC_7
AC_8	AC_9	AC_{10}	AC_{11}	AC_{12}	AC_{13}	AC_{14}	AC_{15}
AC_{16}	AC_{17}	AC_{18}	AC_{19}	AC_{20}	AC_{21}	AC_{22}	AC_{23}
AC_{24}	AC_{25}	AC_{26}	AC_{27}	AC_{28}	AC_{29}	AC_{30}	AC_{31}
AC_{32}	AC_{33}	AC_{34}	AC_{35}	AC_{36}	AC_{37}	AC_{38}	AC_{39}
AC_{40}	AC_{41}	AC_{42}	AC_{43}	AC_{44}	AC_{45}	AC_{46}	AC_{47}
AC_{48}	AC_{49}	AC_{50}	AC_{51}	AC_{52}	AC_{53}	AC_{54}	AC_{55}
AC_{56}	AC_{57}	AC_{58}	AC_{59}	AC_{60}	AC_{61}	AC_{62}	AC_{63}

Figure 1. DCT coefficients (highlighted in grey) used for computation of a text energy

Generally, the symbols in images are almost always larger than 8 by 8 pixels, so the text energy of given symbol is not concentrated inside the block only, but is distributed to its neighbors. In order to achieve better results the average text energy $E_T(i,j)$ of neighbouring blocks instead of the single block is taken in account. The used vertical and horizontal 4 nearest adjacent DCT blocks are shown in Figure 2.

Thus, the text energy $E_T(i,j)$ for DCT block (i,j) is estimated like this:

$$E_T(i,j)=\frac{S_{AC}(i-1,j)+S_{AC}(i,j)+S_{AC}(i+1,j)}{3}+ \\ +\frac{S_{AC}(i,j-1)+S_{AC}(i,j)+S_{AC}(i,j+1)}{3}. \quad (2)$$

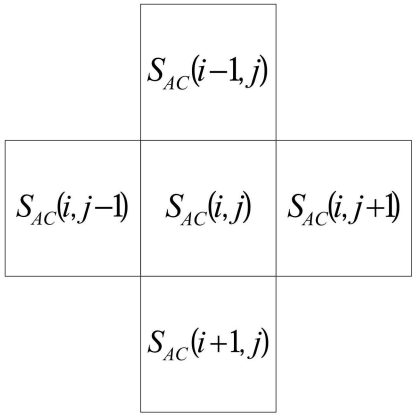


Figure 2. DCT block (i, j) and its neighbors used for computation of the text energy $E_T(i, j)$

In order to determine whether given block belongs to text or background it is necessary to compare each $E_T(i, j)$ to an appropriate threshold value - θ. The image then can be presented as a binary map of white (background / no text) and black (text) blocks - $T_b(i, j)$) , based on the simple decision rule:

$$T_b(i,j) = \begin{cases} 0, & if\ E_T(i,j) > \theta \quad (text) \\ 255, & otherwise \quad (background) \end{cases} \tag{3}$$

The choice of threshold θ should be an image parameters dependent value. In our suggestion θ is based on the average global contrast of image - C . The value of C can be computed easily with sufficient accuracy by finding the difference between the max and min DC coefficient value for the whole image:

$$DC_{\min} = \min\{DC(i,j)\} \tag{4}$$

$$DC_{\max} = \max\{DC(i,j)\} \tag{5}$$

$$C = DC_{\max} - DC_{\min} \tag{6}$$

The optimal relation between C and θ found in our investigations is: θ = 0.9C.

After extraction of text blocks candidates the expectation is each of them to contain a part of text symbols.

Nevertheless, there are some blocks which have quite strong text energy to appear as false alarms. These are edges, noise or other areas in the image that have high AC values. Thereby, additional operations are needed to achieve better extraction results. The following section describes the developed algorithm of 2D filtration operations which is to be applied on the extracted text region candidates. For illustration and describing of the method an example image shown in Figure 3a is processed and each intermediate result is depicted after each step.

First, text blocks are extracted based on the computed DCT text energy (2) and applying the criterion in equation (3). The extracted candidate text DCT blocks can be seen in Figure 3b.

A multi-pass filtration in DCT domain is proposed in the following steps:

1st step: Vertical filtration of single blocks, which are not connected in vertical direction, assuming that text part have height

more than one block - 8x8 pixels (Figure 3c). The operator of the filtration of text DCT block $T_b(i,j) = 0$ utilizes the following criterion:

$$T_b(i,j) = \begin{cases} 255, & if\ T_b(i,j-1) = 255 \\ & and\ T_b(i,j+1) = 255,\ (no\ text) \\ T_b(i,j), & otherwise \end{cases} \quad (7)$$

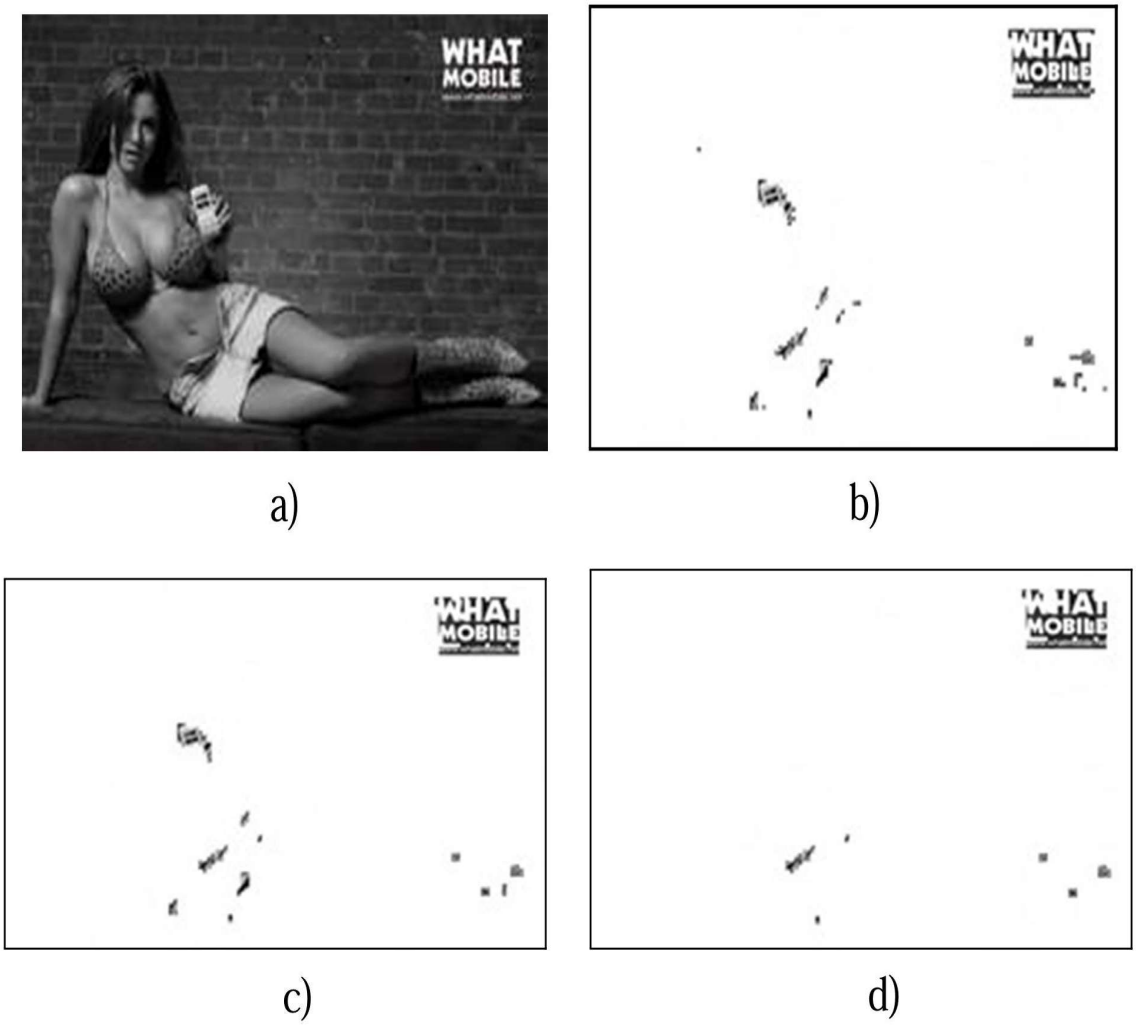
2nd step: Creating rectangle around connected blocks and filtration of vertical structures according to a criterion: , where and are respectively the height and the width of the surrounding rectangle (Fig. 3d).

3rd step: Horizontal filtration of less than four 8x8 blocks, assuming that URL addresses have more than or equal to four horizontal symbols (Figure 3e).

4th step: Vertical filtration of single vertical blocks- second pass (Figure 3f).

5th step: Horizontal filtration of less than four 8x8 blocks - second pass (Figure 3g).

6th step: Creating output image based on the DCT blocks mapped as text (Figure 3h).



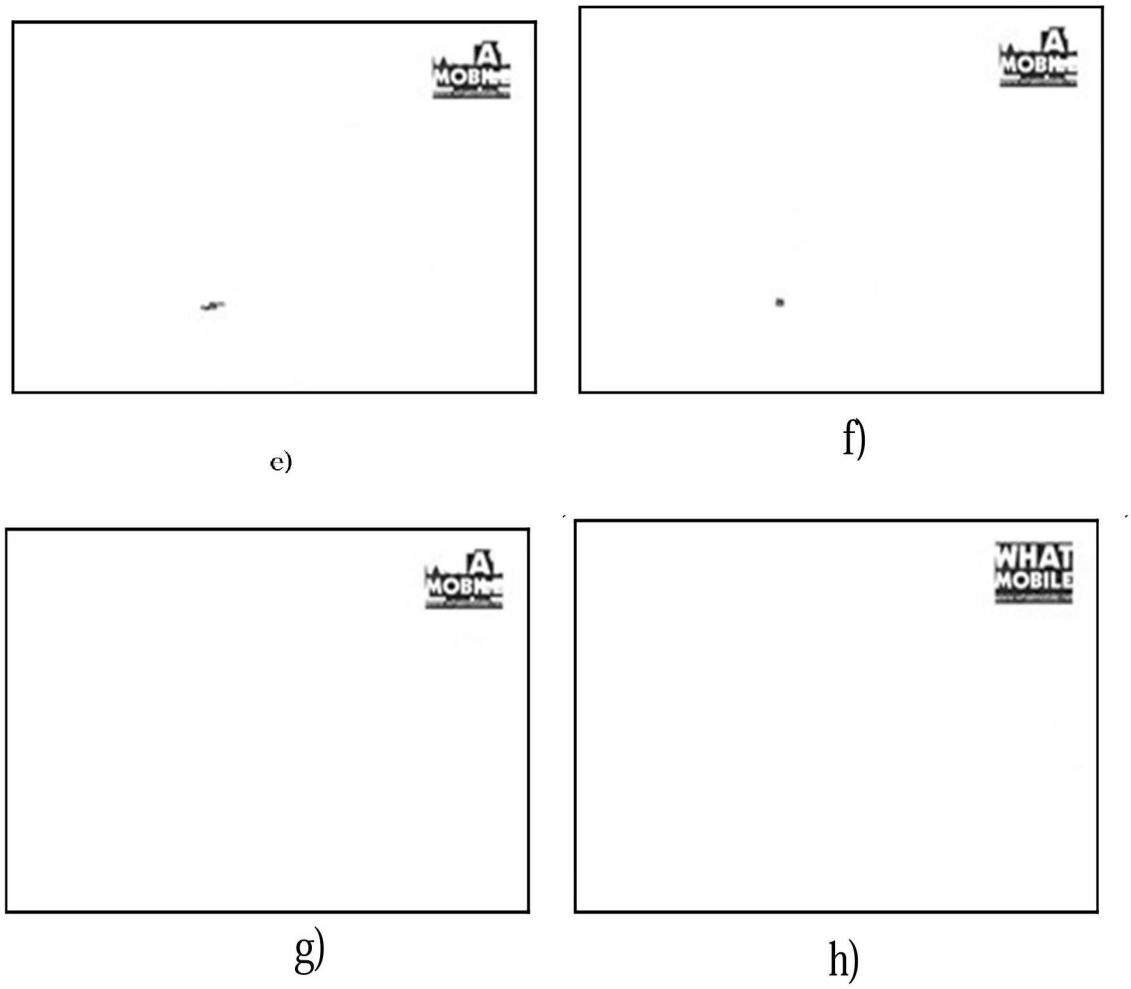


Figure 3. Illustration of intermediate results for text detection and extraction and filtration of an image

In Figure 4 is depicted the developed algorithm block diagram by which the text detection is realized.

3. Experimental Results

The proposed method has been tested on JPEG images with the following characteristics:

Dimensions – from 150 x 220 up to 1900 x 1000 pixels; File sizes – from 7 kB to 444 kB.

Total time for processing 40 test images is 2.26 sec. The average execution time is 0.06 sec/image.

In Table 1 are given the total number of the correctly found text fields, the missed ones and falsely detected ones which don't contain text.

Text region is considered any line of text from one letter to a whole text line filling the image entirely in horizontal direction. Different text lines are considered separate text region regardless of the possibility that they may form one single text block.

To get a general notion of the accuracy of the tested approach two general estimations are given – Recall and Precision, calculated according to (8) and (9) equations respectively:

Correct RegionDetections Missed Regions Recall Correct RegionDetections

$$Recall = \frac{Correct\ Region\ Detections}{Correct\ Region\ Detections + Missed\ Regions}, \quad (8)$$

Correct Region Detections False Alarms Precision Correct Region Detections

$$Precision = \frac{Correct\ \ Region\ \ Detections}{Correct\ \ Region\ \ Detections + False\ \ Alarms} \quad (9)$$

Correct Text Region Detections	Missed Text Regions	False Alarms	Recall, %	Precision, %
98	36	74	90,21	87,36

Table 1. Text Field Accuracy Detection

4. Conclusion

The achieved values according to Table I are high enough corresponding to the current state of similar approaches' performance. The suggested algorithm for text detection in the frequency domain and post-filtration is implemented in the program language C. The result of tests are conducted on an HP® Z600 workstation with dual Quad-core® Intel® Xeon® CPUs, 6 GB DDR2 RAM. The text regions detection accuracy is estimated for images with different dimensions and complex background. The applied 2D filtration in DCT domain decreases the false detected text blocks.

In this paper, we have proposed a combination of DCT features and a new set of schemes for filtration of detected textural, but non-text DCT blocks, differ from previously proposed methods and algorithms, based on the relative DCT coefficients frequency. The method proposed is considered appropriate for the purpose.

Acknowledgement

This paper was supported by the National Science Fund of the Bulgarian Ministry of Education, Youth and Science (Contract – DDVU 02/13 – "Public and Private Multimedia Network Throughput Increase by Creating Methods for Assessment, Control and Traffic Optimization").

References

[1] Crandall, D., S. Antani, R. Kasturi. Extraction of special effects caption text events from digital video. IJDAR(5), No. 2-3, April 2003, pp. 138-157.

[2] Qian, X., G. Liu. Text detection, localization and segmentation in compressed videos. Int. Conf. on Acoustics, Speech, and Signal Processing, *Toulouse, France*, Vol. 2, 2006, pp. II385- II388.

[3] Xu, J., X. Jiang, Y. Wang. *Caption Text Extraction Using DCT Feature in MPEG Compressed Video*. CSIE (6) 2009, pp. 431-434.

[4] Li, H., D. Doermann, O. Kia. *Automatic Text Detection and Tracking in Digital Video*. IP(9), No. 1, January 2000, pp. 147- 156.

[5] Zhong, Y., H. Zhang, A. Jain. Automatic caption localization in compressed video, *IEEE Trans Pattern Anal. Mach. Intell.*, Vol. 22, 2000, pp. 385-392.

[6] Qian, X., G. Liu, H. Wang, R. Su. Text detection, localization, and tracking in compressed video. Signal Processing: *Image*

Communication, Vol. 22, 2007, pp.752-768.

[7] Lu, S., K. Barner. Weighted DCT coefficient based text detection. ICASSP, *IEEE* (2008), pp.1341-1344.

[8] Goto, H. Redefining the DCT-based feature for scene text detection. *Analysis and comparison of spatial frequency-based features*. IJDAR 11(1), 2008, pp. 1-8.

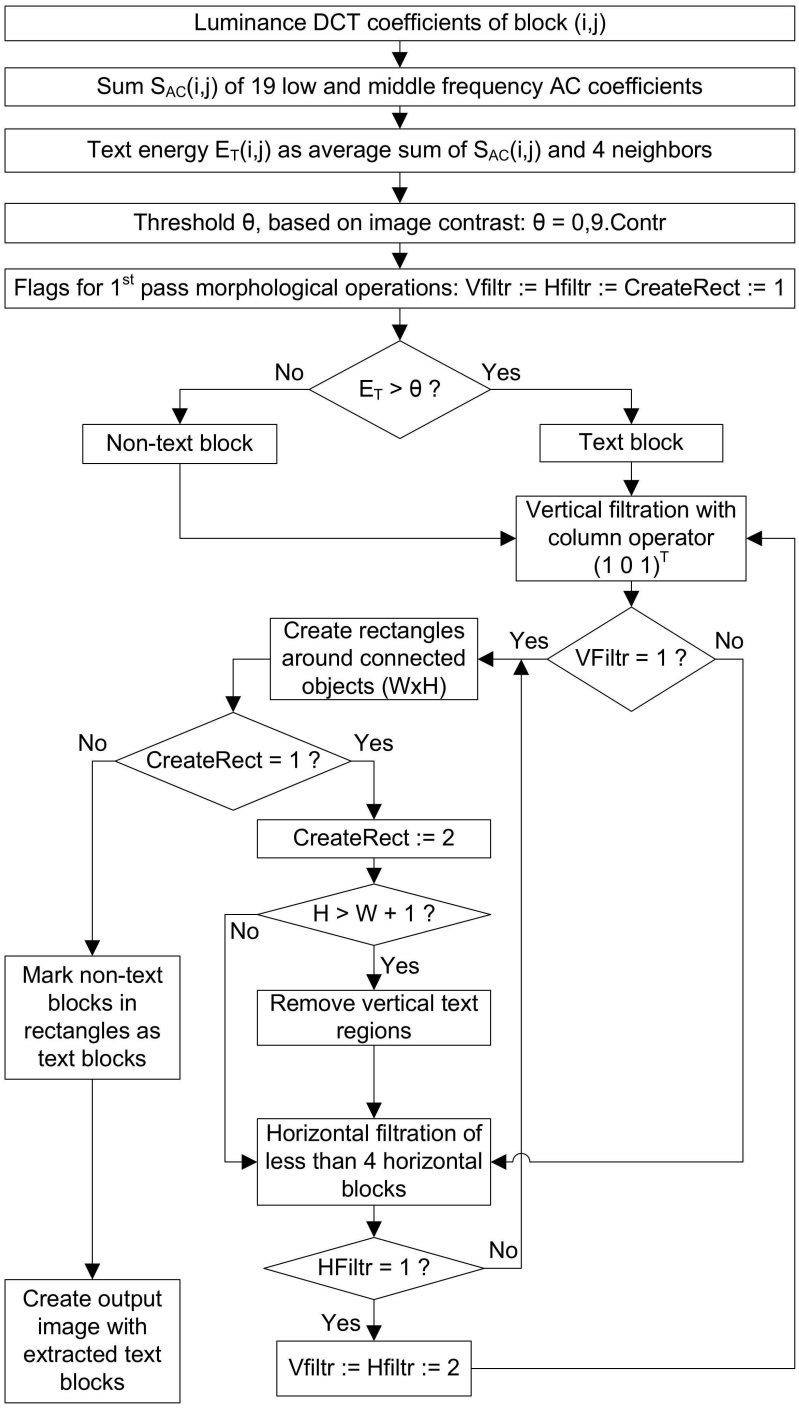


Figure 4. Block diagram of the proposed algorithm for text block extraction