BE- Automatic Text Regions Locating in Digital Images

LIMING CHEN AND JEAN-YVES AULOGE

1. Introduction

In this bureau d'étude, we propose to implement a technique for automatic text regions locating within a digital image as described by an ECL patent by Dr.Walid Mahdi, Dr.Mohsen Ardabilian and Prof. Liming Chen[1]. This method makes use of some simple techniques in image processing, namely multi-resolution, binarization, histogram computation, morphologic operations and intelligent filtering. It was shown on a large dataset of video images selected from various kinds of video programs (commercials, TV news, full-length films, etc.) that this method is capable to locate text regions with different character sizes and styles even in case of complex image background.

These text regions resulted from our technique can then be submitted to an OCR in order to obtain the full texts.

There exist a broad range of applications which may make use of this technique, including still image and video indexing where it may be necessary to extract textual information embedded within images [1] [2] [3] [4], automatic mineral plaque detection and recognition or logo detection and recognition in TV programs.

2. Overview of the Technique

2.1 Characteristics of text regions within a digital image

Textual information in a digital image such as a TV image can be classified into two kinds: *natural text* which appears as a part of the scene (e.g. street names or shop names in the scene), and *artificial text* which is produced separately from the video shooting or image capture and inserted into the scene during a post-processing step, for instance by a video title machine. Both of them, when they occur within a digital image, are of capital impotence and they are good clues for content-based indexing and retrieval. However, by the opposition to the *natural text* which accidentally appear in the scene, the inclusion of artificial text is carefully selected, and thus is subjected to many constraints so that the artificial text is easily read by viewers. Below we summarize the main characteristics of these constraints by the following features [7]:

- Text character are in the foreground
- Text characters contrast with background since artificial text is designed to be read easily.
- Text characters are monochrome.
- Text character has size restrictions. A letter is never as large as the whole screen.
- Character size should not be smaller then a certain number of pixels otherwise they are illegible to viewers.

2.2 Flowchart of the text regions locating process

The proposed method makes use of these basic features to localize text regions in images. It also takes into account the characteristics of digital images with complex background, for instance video images, such as the low resolution, presence of noises, the absence of control parameters in the cameras. Actually, the majority of text region detection methods in the literature tend to put up with the hallucination of some text regions which is more acceptable than missing the detection.

The proposed technique is composed of the following steps :

1. Digital image transformation ;
2. Enhancement of Text region patterns
   A - Multi-resolution processing ;
   B - Binary operation for text region enhancement ;
3. Potential text regions detection ;
4. Effective text regions selection ;

[1] W.Mahdi, M.Ardebilian, L.Chen, "Localisation de textes dans les images", PCT/FR03/02406, 31 Juillet 2002

It should be noted that steps 2.A and 2.B are permutable. Figure.1 summarizes the steps of our approach. In the following, we describe in detail each of theses steps.

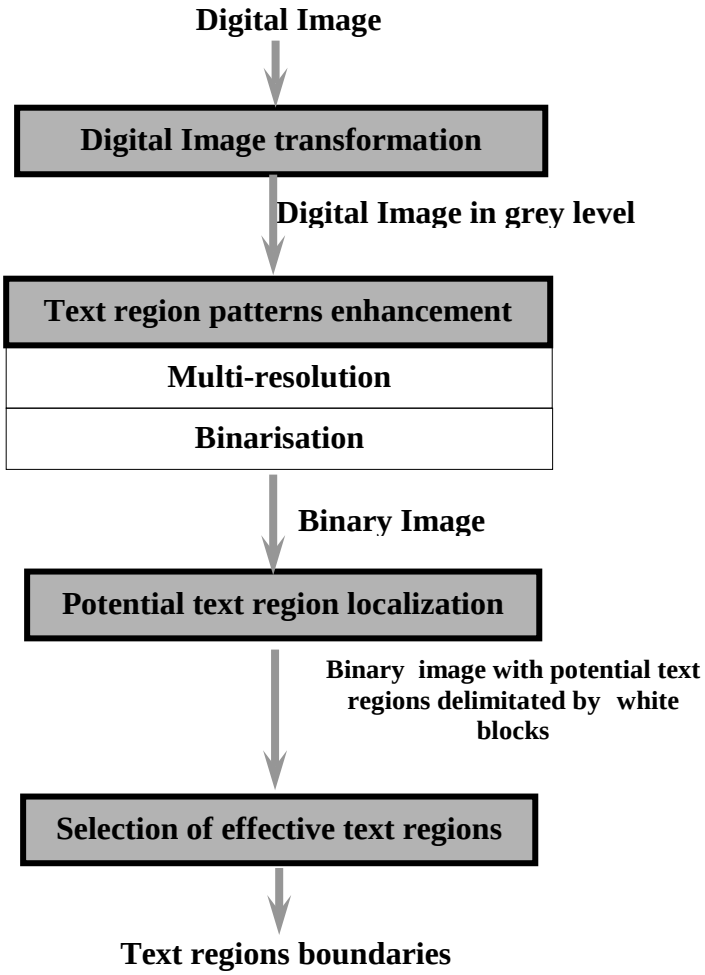


Figure 1 - Overview of text regions locating process

3. Implementation

In this bureau d'étude, we propose to implement step by step the text locating technique as described in the patent.

3.1 Digital image transformation

We assume that the input image is in bitmap format. When the input image is in a compressed format such as JPEG for instance, we first convert it into bitmap format.

This step consists of the following stages:

1. If the input digital image is a colour image, say I, convert it into an image in grey level, say G ;
2. If necessary, compute the transpose of G ;
3. If possible, define a sub-zone of the image in which text regions are going to be looked for.

Stage 1 is a simple conversion of a digital colour image, generally represented by three matrix in a colour space, RGB for instance, into a one in grey levels. This stage is necessary and indispensable for the stage of binarisation which will be described later.

The purpose of stage 2 is to make possible the detection of the regions of vertical texts. To be able to discover regions of vertical texts, we can either first transpose an input digital image, represented by a matrix I, to obtain a new image represented by ${}^{t}I$, and afterward apply to it a set of morphological operators $\{ M_n \}$, or apply the set of transposed morphological operators $\{ {}^{t}M_n \}$ directly to the input image I. Formally, these two processes are identical.

Stage 3 aims at accelerating the method by reducing the zone of search for text, for example, in case we are sure that texts appear in a fixed region of the image.

3.2 Enhancement of text region patterns

The localization of probable text regions within an image generally is a part of preprocessing which is very important for text detection purpose.

The proposed technique first makes use of a multi-resolution approach and conversion of a grey level image into binary image to enhance probable text regions. Notice that these two stages are interchangeable.

The conversion of an input grey level image I into a binary image BW consists of thresholding. That is, the output binary image BW has value of 0 (black) for all pixels in the input image I with value less than a predefined threshold and 1 (white) for all other pixels.

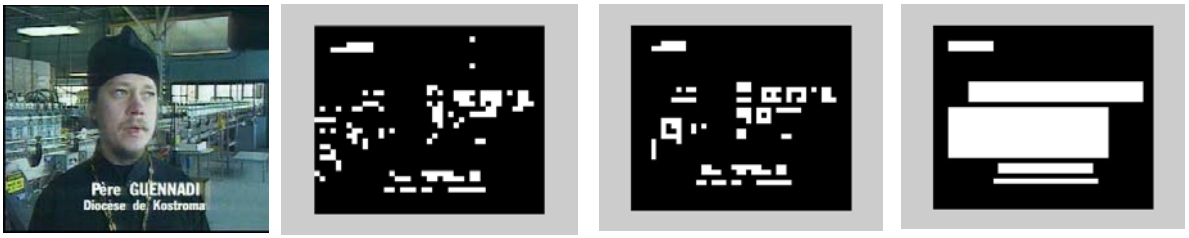
The use of the multi-resolution method for text lines localization is based on the basic feature that a text line generally appears in the shape of a full line in an image with low resolution.

The multi-resolution process, when applied to an input image I, returns an output image J that is M times the size of I. If M is between 0 and 1.0, image J is smaller than I. If M is greater than 1.0, J is larger than I. This passage from an initial resolution of an image I to a new resolution J is ensured by a specified interpolation method.

In the proposed technique, M was set to 0.125 and the nearest neighbor interpolation method was used. Figure 2.b shows the results of such a process. Notice that other interpolation techniques, such as linear interpolation, may also be used.

It should be noted that the parameter M may be changed, for example, according to the size of the image. The proposed text region locating technique does not depend on the value of this parameter as long as this one is between 0 and 1. The value of the threshold used for converting an input grey level image into a binary image may be changed as well, for instance according to the type of the input image. In the following, we propose to set it to 0.7.

The examination of figure 2.b clearly shows that the multi-resolution method makes it possible to filter the input image and to only keep connected components with homogenous color corresponding to a meaningful area.



-a- *-b-* *-c-* *-d-*

Figure.2 Probable text region localization process. a) an input digital image with complex background. b) an output binary image after text region patterns enhancement.. c) the output image "*b*" after negative form elimination. d) Image "*c*" after potential text regions localization.

3.3 Potential text regions localization

Once obtained a binary image with text region patterns enhanced, this step aims at localizing potential text region within an image.

It consists of applying morphological masks to obtain the closure of the blocks susceptible to contain texts by filling empty zones between characters or words.

Algorithm: **Potential text regions localization**
Input : a binary image I ;
Output : a binary image J depicting potential text regions by rectangle white blocks ;
Begin
Repeat
Apply binary morphological operations *;*
Until (the image has no large changes)
End.

Figure 2.d. illustrates the result of this algorithm applied to figure 2.c.

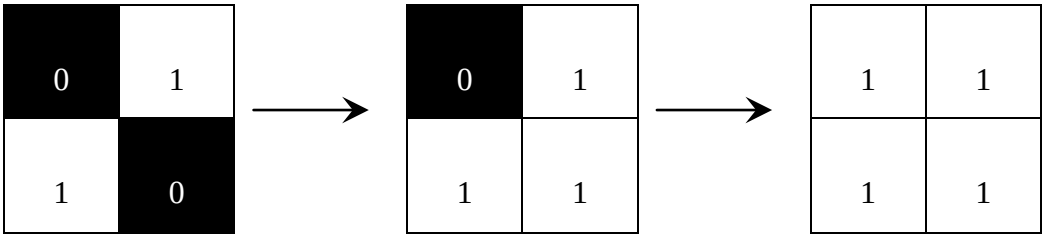
In the following, are given three different morphological masks which may be used for the closure of blocks susceptible to contain texts. All these masks may be combined and applied in different order. Whatever the morphological masks used or the order in which they are applied, the important here is to obtain the closure of blocks susceptible to contain texts.

The first morphological mask is depicted by figure 3. As we can see in the figure, we set all pixels to 1 when the border pixels on the left and on the right are valued by 1.

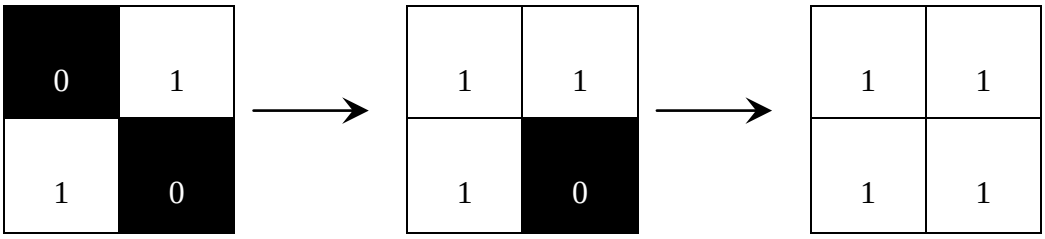
The second morphological mask is depicted by figure 4. As we can see in the figure, this mask leads to a diagonal closure when two border pixels on the diagonal are valued by 1.

The third morphological mask is depicted by figure 5. As we can see in the figure, this mask is similar to the previous one. It aims at a diagonal closure as well.

Below we illustrate two cases which are applications of the third morphological mask M_3. The first case is a two step applications of M_3, by setting to 1 first the 0 on the bottom right position, then the 0 on the up left



position, leading to the final square of all pixels set to 1. The second case is symetrical to the previous one while first setting to 1 the 0 on the up-left position. Notice that we can also proceed in parallel as sugested by figure 5.



The detection of text regions on the original input image can be derived from the mapping between coordinates of text regions detected in the binary image and coordinates in the input image.

Figure 2.d shows all probable text regions obtained after applying such binary operations. However, we can notice that figure 2.d contains five areas candidates of text regions whereas on the source image there are only two effective text regions. The next step tackles this problem.

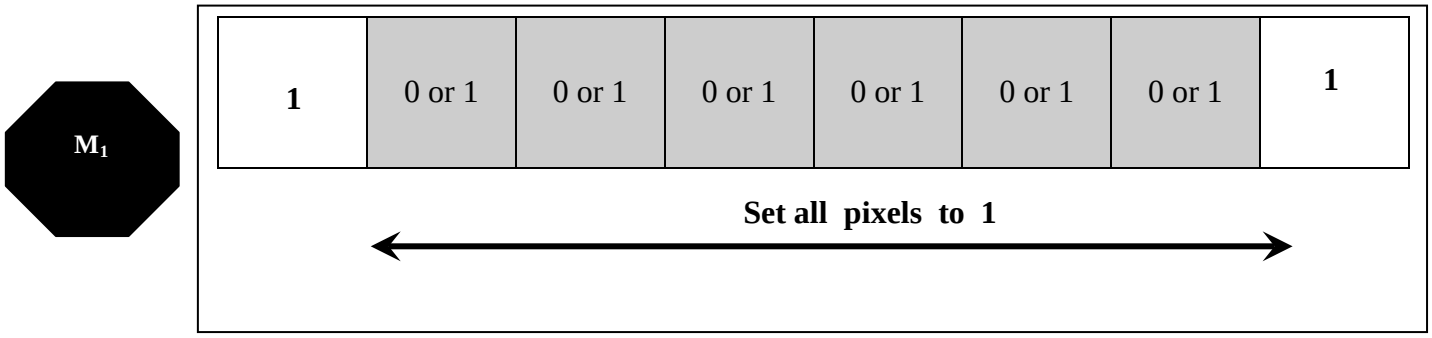


Figure 2 - Morphological mask M_1

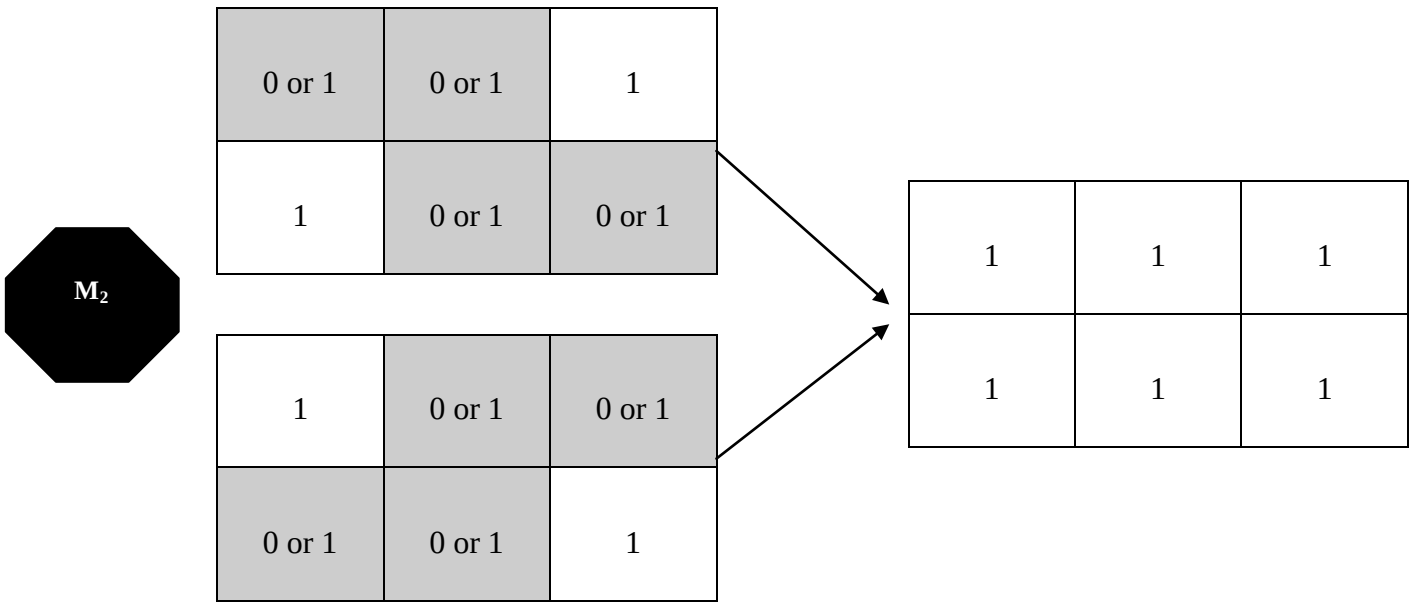


Figure 3 - Morphological mask M_2

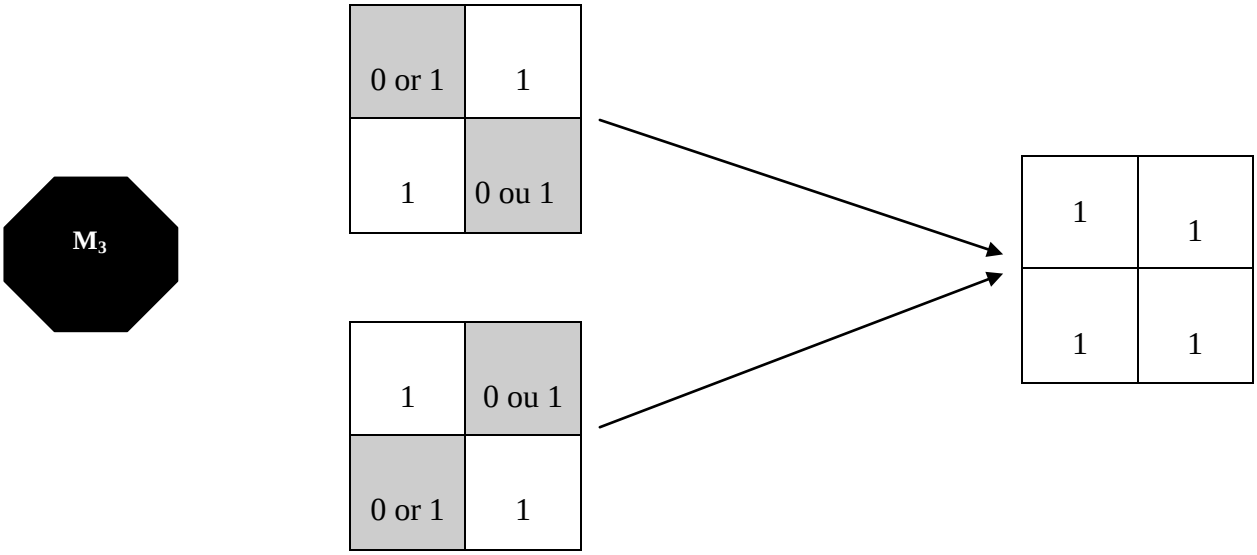


Figure 4 - Mophological mask M_3

3.4 Selection of effective text regions

The presence of the probable text regions in figure 2.d which actually are not text regions can be explained by the fact that the multi-resolution method based on Bi-level (binary image) thresholding is a method rather efficient for textual document applications for which a pixel belongs to the background or to the certain image object. In the case of a digital image with complex background such as video images, an image generally is made of several objects with different colors. Consequently it is possible to have some false detection of text regions when applying multi-resolution method based on Bi-level thresholding to such kind of images.

In the proposed approach the results obtained from the previous step constitute a first localization of candidate regions susceptible to contain a text. We actually perform a further step which examines each candidate region which has been localized in order to decide whether it is an effective text region or not. This selection of effective text regions is a two stage process which consists of background pixels separation and effective text region filtering.

3.4.1 Background pixels separation

The stage of background pixels separation aims at highlighting character pixels from the background. For this purpose we apply an intensity level slicing method to the gray level image obtained after the first image transformation step. This technique is useful when different features of an image are contained in different gray level. It consist of mapping each gray level a $\in$ [0, L] into a gray level v $\in$ [u, L] according to a transformation depicted by equation 1 :

$$v = f(a) \qquad (1)$$

which can be simply defined by Equation 2 :

$$v=\begin{cases} a, & a\leq u \\ L, & otherwise \end{cases} \qquad (2)$$

Another transformation possible for the purpose of highlighting character pixels can be defined by Equation 3 :

$$v=\begin{cases} u, & a\leq u \\ L, & otherwise \end{cases} \qquad (3)$$

Furthermore, the interval [u, L] is determined dynamically from the histogram " H " of the grey level image J obtained from the input image after transformation process. Recall that image J may potentially have 256 different grey levels. The process for choosing dynamically the interval [u,L] is the following

1. L is initialised with the value 255 which represents the white colour ;
2. To determine the value of u, one begins by calculating number of pixels ***Nb*** having the value 255, then one adds gradually to ***Nb*** number of pixel having the colour 254, 253 , 252, etc.; until ***Nb*** is superior to 2 % of the total number of pixels of the whole grey level image J. The last value of the histogram H, taken into account in this operation is then affected to u.

Of course, the threshold 2% of the total number of pixels used in our experiment may need to be adapted according to applications.

3.4.2 Effective text region filtering

Now a simple analysis of the spatial variation of all candidate text regions, transformed by the previous background pixels separation, allows us to identify effective text regions.

This analysis is based on a basic text characteristics which says that text characters generally contrast with background since artificial text is designed to be read easily. All we need to do is to locate the two most important peaks (local maximum) in the histogram of each transformed potential text region in order to obtain their positions P1 and P2. Figure 5 illustrates such a process applied to the potential text regions identified by Fig.5-a, which results from the potential text regions localization step on the grey level image Fig.2-a.

A spatial variation of each candidate text region is then characterized by equation 4.

$$D(P_1, P_2) = abs\ (P_1 - P_2) \qquad (4)$$

If the distance D(P1, P2) is greater than a predefined threshold, the candidate text region is classified as an effective text region, otherwise it will be ignored.

As we can see in figure 5, it is clear that regions 1, 2 and 3 have a weak spatial variation. Consequently these regions are ignored. The first results are detected with a threshold value equal to 15% of the total bin number in the gray scale levels. Higher is the threshold value and better is the method accuracy.

3.5 Improvements

Some improvements can be made to this text region locating process in order to better capture the text region boundaries or speed up the whole process by eliminating some evident negative potential text regions.

3.5.1 Improving text region localization

This stage aims at the capture of the complete text area by recursively applying background pixels separation process to each effective text region already transformed. Indeed, the text regions obtained by all the treatments presented previously may lead to a non complete text area which is possibly caused by the interpolation method used in the multi-resolution process. To obtain the final results which give the horizontal and vertical boundaries for each text region, we need a further processing.

3.5.1.1 Horizontal delimitation of text region boundaries

We first select a representative horizontal line $Rh_{lg}(i)$ among all lines within a text region which has been identified by the process.

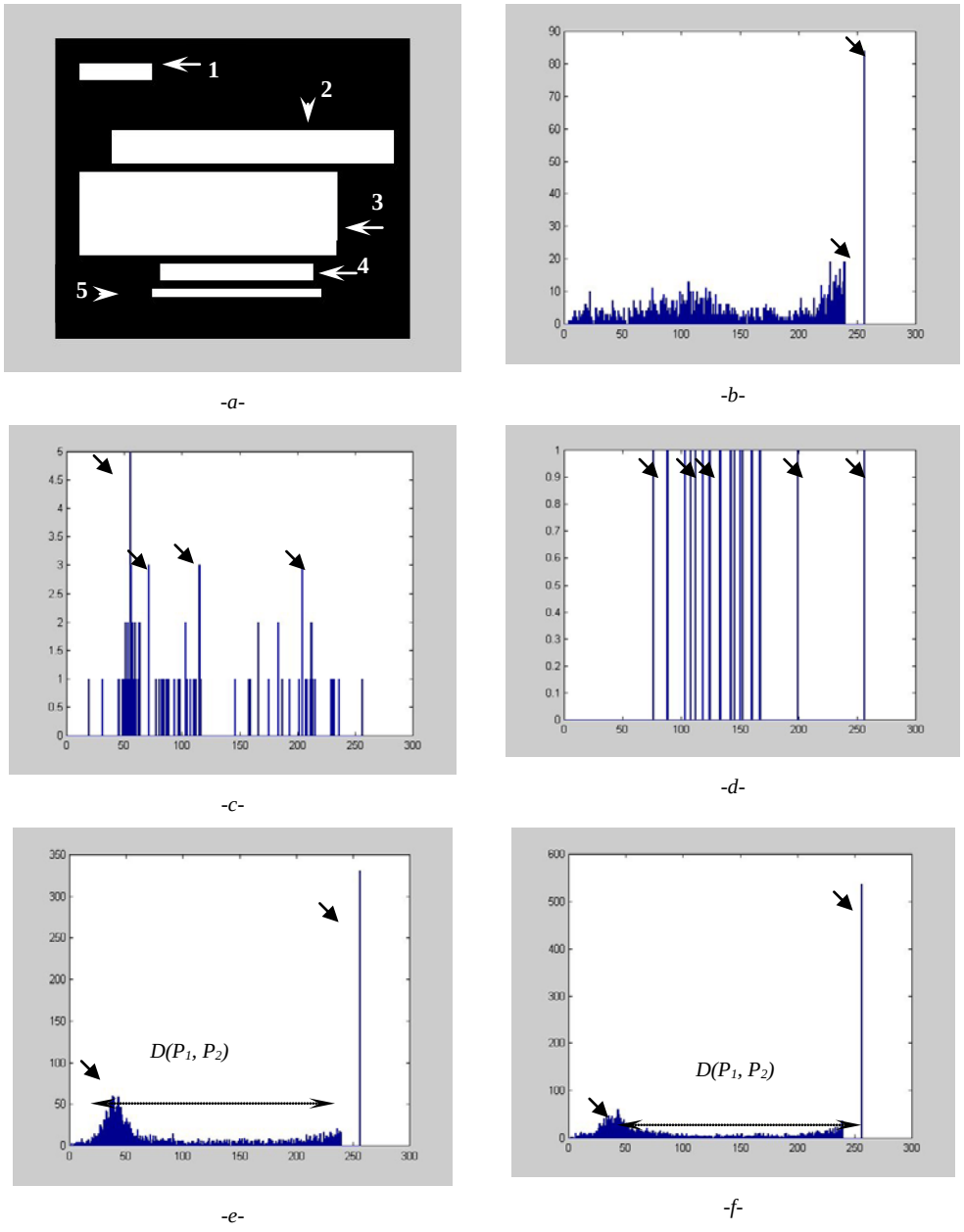
For the choice of $Rh_{lg}(i)$ we propose to select the one which is formed by the maximum of pixels horizontally aligned and belonging to characters. Generally the selected line $Rh_{lg}(i)$ will be the one formed by the maximum pixel number having a value equal to L since after background pixel separation transformation, characters in a text region are assumed to be monochrome and contrasting with their background.

Next, we compare $Rh_{lg}(i)$ with adjacent line $Rh_{lg}(i-1)$ that immediately precedes it (respectively follows it $Rh_{lg}(i+1)$) in order to decide to merge the two lines into one text block or not. For this purpose, the merging criterion is based on the spatial grey value distribution and principle of connected monochrome pixels as follows :

Let $Pos_{Rhlg(i)}$ and $Pos_{Rhlg(i-1)}$ (respectively $Pos_{Rhlg(i+1)}$) to be two sets that describe pixel positions in line $Rh_{lg}(i)$ and $Rh_{lg}(i-1)$ (respectively $Pos_{Rhlg(i+1)}$) , having grey value equal to *L*. Now consider the equation

$$Pos_{Rhlg(i)} \cap Pos_{Rhlg(i+1)} \neq \varnothing \qquad (5)$$

If equation 5 is satisfied, $Rhlg(i)$ is replaced by $Rh_{lg}(i-1)$ (respectively $Pos_{Rhlg(i+1)}$), and the process is recursively applied until the complete stabilization of the bottom and up horizontal boundaries of the text region.



-a- *-b-* *-c-* *-d-* *-e-* *-f-*

Figure 5 - Histogram of candidate text regions after applying background pixels separation. Figure b, c, d, e and f respectively illustrate the histogram of region 1, 2, 3, 4 and 5 of figure a.

3.5.1.2 Vertical delimitation of texts region boundaries

For the vertical delimitation of text region boundaries, we propose to add to the representative line $Rh_{lg}(i)$ all pixels which satisfy the following conditions :

- Only pixels which are on the left or on the right of pixels forming the representative line $Rh_{lg}(i)$ are considered.
- Only pixels having the same color value than $Rh_{lg}(i)$ pixels, are added to $Rh_{lg}(i)$.
- Pixels appended to the $Rh_{lg}(i)$ line must respect the negative form elimination principle presented in next section.

3.5.1.3 Interchanging horizontal and Vertical delimitation of texts region boundaries

For the detection of text regions in vertical position within a digital image, it may be necessary to apply the principle presented in 2.2.5.1.1 Horizontal delimitation of text region boundaries first in a vertical way. However, this case amounts to apply a transposition to the input image as we present in 2.2.1 Digital image transformation.

3.5.2 Negative form elimination

We can speed up the text region locating process if we have some specific knowledge on possible text regions. One of these techniques that we can apply for video images for instance is the so called *negative form elimination* which consist of eliminating all connected components of homogenous color belonging to the edge pixel of an image and all horizontal lines which are longer than a predefined threshold *lt* .

The last elimination operator can be represented by the morphological operator M4 illustrated by figure 6. The basic features of text regions characteristics in video images that we have outlined in 2.1 Characteristics of text regions within a digital image, justify well such a processing. Figure 2-c illustrates the result obtained once the negative form elimination has been applied to figure 2-b.

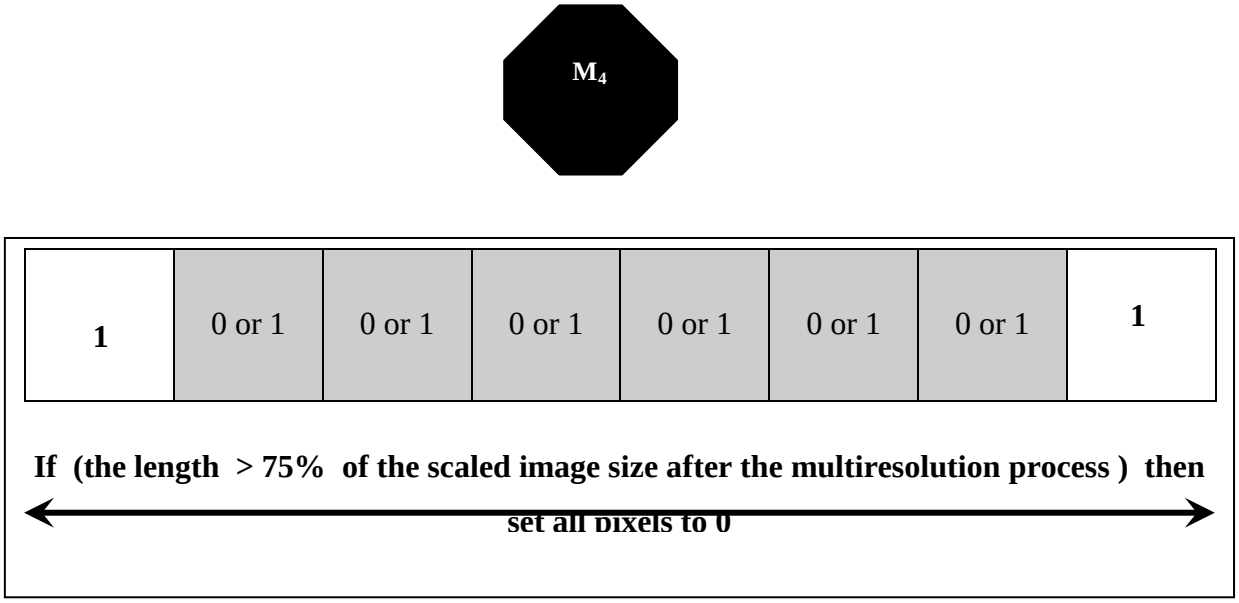


Figure 6 - morphological operation M_4

Another improvement is the diagonal fill to eliminate 8-connectivity of background as illustrated by the morphological mask M_5 in figure 7. This operator consists of setting to 0 an isolated pixel valued by 1 while it is surrounded by 0 pixels.

When this operator M_5 is applied before the morphological operators M_1, M_2 and M_3 we increase the precision of the text region boundaries detected in the image by eliminating isolated pixels which may be merged to effective text regions while applying morphological operator M_1.

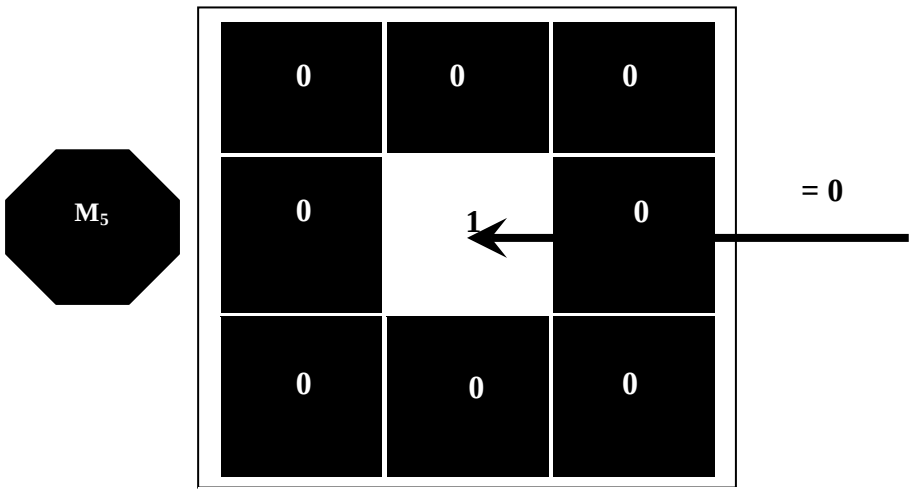


Figure 7 - Morphological operator M_5

4. Experiments

In order to experiment the proposed technique, we propose to experiment your implementation on a set of digital images extracted from different film genres : commercials, newcasts and feature films compressed in *MPEG1*. Please show the located text regions on each image.

5. Questions

5.1 How should we evaluate the accuracy of the proposed technique ? what criteria should be used ?

5.2 What is the potential drawbacks of the proposed technique. Please motivate your answer.

Figure 8. Experimental results of the proposed method applied to video images with complex background

6. References

[1] W Mahdi, M. Ardebilian, L Chen ., «Automatic Video Scene Segmentation based on Spatial-temporal Clues and Rhythm», International Journal of Networking and Information Sytems. Special Issue on Video Data, Vol N°3, décembre 2001.

[2] W Mahdi «Macro-segmentation sémantique des documents audiovisuels à l'aide des indices spatio-temporels », Thèse de doctorat sous la direction de L.Chen, Ecole centrale de Lyon, 2001.

[3] M. Ardebilian, L.Chen et X.W.Tu, « Robust Smart 3-D Clues Based Video Segmentation for Video Indexing », Journal of Visual Communication and Image Representation, Volume 11(1), Mars 2000.

[4] Y. Chahir, « Indexation et recherche par le contenu d'informations visuelles », Thèse de doctorat sous la direction de L.Chen, Ecole centrale de Lyon, 1999.

[5] Zhong, Y., Kary, K., jain, A.K, " Locating text in complex color Images", pattern recognation, Vol.28, N° 10, 1995, page 1523-1535.

[6] Ohya, J., Shio A., Akamatsu, S., "Recognizing characters in scene images", IEEE Trans. On PAMI, Vol 16, N° 2, page 214-220, February 1994.

[7] W. MAHDI, M. ARDEBILIAN, L. CHEN, "Automatic Text Regions Locating for Smart Video Access", The Third International Conference on Computer-AIDED Industrial Design and Computer-AIDED Conceptual Design, IAP Ed., ISBN.7-5062-2011-3. HONG KONG, Polytechnic University. 26-28 Novembre 2000, pp.380-388.