TEXT LOCALIZATION IN COLOR DOCUMENTS

N. Nikolaou*, E. Badekas*, N. Papamarkos* and C. Strouthopoulos#
*Image Processing and Multimedia Laboratory, Department of Electrical & Computer Engineering
Democritus University of Thrace, 67100 Xanthi, Greece,
Department of Informatics and Communications, Technological Educational Institution of Serres
62123 Serres, Greece,

Keywords: Document processing, Text localization, Page layout analysis, Color quantization.

Abstract. A new method for text localization in cover color pages and general color document images is presented. The colors of the document image are reduced to a small number using a color reduction technique based on a Kohonen Self Organized Map (KSOM) neural network. Each color defines a color plane in which the connected components (CCs) are extracted. In each color plane a CC filtering procedure is applied which is followed by a local grouping procedure. At the end of this stage, groups of CCs are constructed which are next refined by obtaining the Direction Of Connection (DOC) property for each CC. Using the DOC property, the groups of CCs are classified as text or non text regions. Finally, text regions identified in the different color planes are superimposed and the final text localization of the entire document is achieved. The proposed technique was extensively tested with a large number of color documents.

1 INTRODUCTION

Interest about exploiting text information in images and video has grown notably during the past years. The ability of text to provide powerful description of the image content, the convenience of distinguishing it from other image features and the provision of extremely important information, reasonably attracts the research interest. Content-based image retrieval, OCR, page segmentation, license plate location, address block location and compression are some examples based on text information extraction from various types of images.

A main categorization of text localization methods include texture based techniques (Jain and Zhong, 1996), (Jain and Bhattacharjee, 1992) and connected components (CCs) based techniques (Strouthopoulos et al., 2002), (Chen and Chen, 1998), (Sobottka et al., 2000), (Hase et al., 2001), (O'Gorman, 1993) and (Fletcher and Kasturi, 1988). Some hybrid approaches have also been reported in the literature (Zhong et al., 1995) and (Jung and Han, 2004). Texture based techniques are time consuming and use character size restrictions. The main advantage is their capability of detecting text in low resolution images. On the other hand, CCs based techniques are fast and exploit the fact that characters are segmented.

This paper was partially supported by the project Archimedes 1, TEI Serron.

Most approaches for text localization refer to gray or binary document images. Only recently, some techniques have been proposed for text localization and extraction in color documents. Strouthopoulos et al. (2002) proposed a method for text extraction in complex color documents. It is based on a combination of an adaptive color reduction technique and a page layout analysis approach which uses a neural network block classifier in order to identify text blocks. Zhong et al. (1995) presented a hybrid system for text localization in complex color images. According to this system, a color segmentation stage is performed by identifying local maxima in the color histogram. Heuristic filters on the CCs of the same color plane are applied and non-character components are removed. A second approach based on local spatial variance, which locates text lines, is also proposed. Chen and Chen (1998) proposed a method for text block localization on color technical journals cover images. Initially, the colors of the image are reduced using a YIQ color model based algorithm. With the Sobel operator and through a binarization process, strong edges are isolated. Primary blocks are then detected with the Run Length Smearing Algorithm and finally classified with the use of nine features which underlie on fuzzy rules. Sobottka et al. (2000) proposed an approach to extract text from colored

Nikolaou N., Badekas E., Papamarkos N. and Strouthopoulos C. (2006).
TEXT LOCALIZATION IN COLOR DOCUMENTS.
In *Proceedings of the First International Conference on Computer Vision Theory and Applications*, pages 181-188
DOI: 10.5220/0001365801810188

books and journal covers. The image is quantized with an unsupervised clustering method and the text regions are then identified combining a top-down and a bottom-up technique. An algorithm for character string extraction from color documents is presented by Hase et al. (2001). First the number of representative colors of a document is determined. Potential character strings are then extracted from each color plane using multi-stage relaxation. When all extracted elements are superimposed, a strategy which utilizes the likelihood of a character string and a conflict resolution is followed to produce the final result. A detailed review on text information extraction techniques is presented by Jung et al. (2004).

As it is mentioned above, in most of the cases a localization technique for complex color documents includes a color quantization stage. The perfect reduction of the document colors is crucial for the effectiveness of the entire localization technique. The goal of this paper is to propose a new technique for text localization which can overcomes the difficulties associated with mixed type color documents such as complex color cover pages. Specifically, for this type of color documents, text and graphics are highly mixed with the background and even more, in many cases, the background cannot be defined. The proposed technique efficiently integrates a KSOM based color quantization procedure and a color plane analysis technique. That is, in order to handle varying colors of the text in the image, a combination strategy is implemented among the binary images (we call them color planes) obtained by the color quantization. Specifically, after color quantization, each color defines a color plane in which the connected components (CCs) are extracted and a CC filtering procedure is applied which is followed by a local grouping procedure. At the end of this stage, groups of CCs are constructed which are next refined by obtaining the Direction Of Connection (DOC) property for each CC. Next, the groups of CCs are classified as text or non text regions using their DOC property. Finally, text regions identified in the different color planes are superimposed and the final text localization of the entire document is achieved.

The proposed text localization technique consists of the following main steps.

Step 1. Automatic estimation of the document image dominant colors.
Step 2. Color reduction by using the KSOM and color planes extraction.
Step 3. Connected component filtering
Step 4. Initial elements grouping
Step 5. Final elements grouping and blocks identification
Step 6. Classification of blocks
Step 7. Color planes superimposition and final color text localization.

Steps 3-6 are applied separately on each color plane produced by step 1. The grouping procedure of CCs into homogenous sets (steps 4 and 5) is carried out in two phases. First we create connections between CCs (component pairing) by searching for similar objects in an adaptively defined area. This will lead to a general formation of groups. Based on features resulted from the connections, we assign a special property to each CC named Direction Of Connection (DOC) which indicates whether a CC is likely to belong in an horizontal or a vertical structure. This information is used to remove false connections between objects and perform the final forming of element groups. Classification module labels the groups as text or non text and finally the results are adopted from the last step which superimpositions the detected text regions from all color planes.

The method performs satisfactory in the majority of mixed type of color documents. However, it is preferable the document to satisfy the following conditions that are satisfied by the majority of modern book covers and generally color documents:

- The color of the characters should not be gradient.
- Text orientation is allowed to be horizontal and/or vertical with about 15 degrees of angle tolerance.

The proposed technique is implemented in visual environment and it has been extensively tested with success with a large number of color documents.

2 COLOR QUANTIZATION

In the first step of the proposed technique, the number of dominant colors appeared in the color document is estimated and then the final dominant colors are obtained. The estimation of the number of dominant colors is performed by identifying the main peaks of the image luminance histogram (Atsalakis, 2002). The dominant colors are obtained by using the method for color reduction proposed by Papamarkos (1999) which is based on a KSOM neural network. According to this method, a KSOM neural network is fed by the image colors and the

output neurons define, after training, the centers of the color classes obtained.

Usually, the number of dominant colors is not greater than eight. This estimation corresponds to a smaller number of colors than the ideal case but this prevents the oversegmentation of the characters. Also, in our tests very rarely this resulted in fusion of characters with the background due to the large color distance between characters and background.

3 CONNECTED COMPONENTS PROPERTIES AND FILTERING

In each color plain, connected components (CCs) are identified. The enclosing rectangle of a connected component (CC) is defined as its bounding box. Let CC_i be a connected component. Every CC is characterized by the following set of features:

i. $BB(CC_i) = \{W_i, H_i\}$. The bounding box of CC_i. W_i represents the width and H_i the height.
ii. $psize(CC_i)$. The pixels count of CC_i.
iii. $bsize(CC_i) = W_i \times H_i$. The size of $BB(CC_i)$.
iv. $dens(CC_i) = psize(CC_i) / bsize(CC_i)$. The density (or saturation) of CC_i.
v. $elong(CC_i) = \min\{W_i, H_i\} / \max\{W_i, H_i\}$. The elongation of CC_i.

According to the above features, CC_i is considered as a non text object if

i. $psize(CC_i) < T_{psize}$, where T_{psize} is taken equal to 6 pixels.
ii. $dens(CC_i) < T_{dens}$, where T_{dens} was set at 0.08. CC_i must cover no less than the 8% of the $BB(CC_i)$.
iii. $elong(CC_i) < T_{elong}$, where T_{elong} was set at 0.08. This means that the width W_i of a CC cannot be 12.5 times larger than H_i (and the opposite).

Thresholds have been carefully selected after several tests in order not to reject character elements. This filtering procedure can be considered as a preprocessing step and targets only on removing very noisy elements resulted from the color reduction procedure. In addition, it speeds up the entire text localization technique since the number of CCs decreases significantly.

4 BLOCK FORMATION

4.1 Initial Grouping of CCs

Let CC_i be a connected component. We define a region $R(CC_i)$ (Fig. 1(a)) as the set of all pixels (x_i, y_i) satisfying the following inequality:

$$d_{\min} \le d_i \le d_{\max} \quad (1)$$

where d_i the Euclidean distance of pixel (x_i, y_i) from the centroid of CC_i, $d_{\max} = c_d \max\{W_i, H_i\}$ and $d_{\min}$ a small constant value (usually 5 pixels). Coefficient c_d adjusts the size of $R(CC_i)$ and if it is taken equal to 3 then the neighborhood region defined contains the necessary information that will specify if the CC_i belongs to an horizontal or vertical structure block.

Only objects whose centroid is located inside $R(CC_i)$ are considered as objects that it is possible to be connected with CC_i. For a connection to be established between a pair of CCs (CC_i, CC_j), the following similarity criterion, based on *psize*, must also stand.

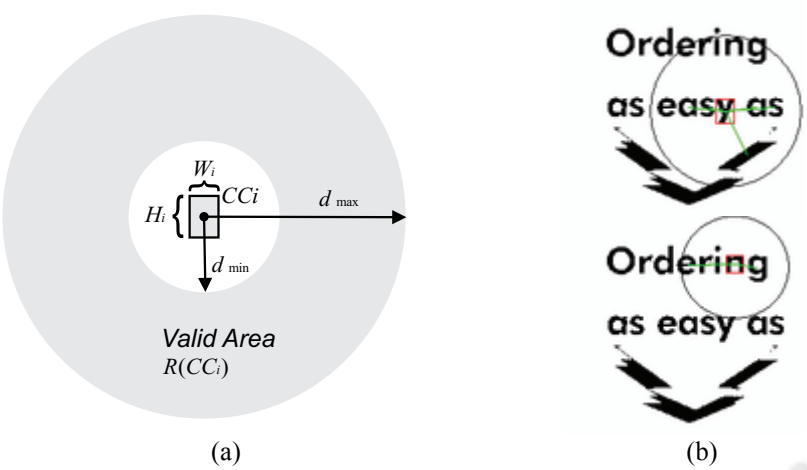
$$\frac{\max\{psize(CC_i), psize(CC_j)\}}{\min\{psize(CC_i), psize(CC_j)\}} \le T_{psr} \quad (2)$$

T_{psr} takes the large value 7 in order CCs having *psize* ratio larger than this threshold value not to be connected. Tests that were contacted on clean text images of various fonts showed that character elements have size ratio differences no more than 6.5 even in the case of joined characters, so the value of 7 was chosen to have an additional safety tolerance. Fig. 2 shows such an example.

The final constraint for CCs connection is related to a distance measure defined in (Simon et al., 1997). This distance measure is adopted but it is used in a different way. We define that for the two CC_i and CC_j the Horizontal Block Distance ($HBD(i, j)$) and the Vertical Block Distance ($VBD(i, j)$) between them is

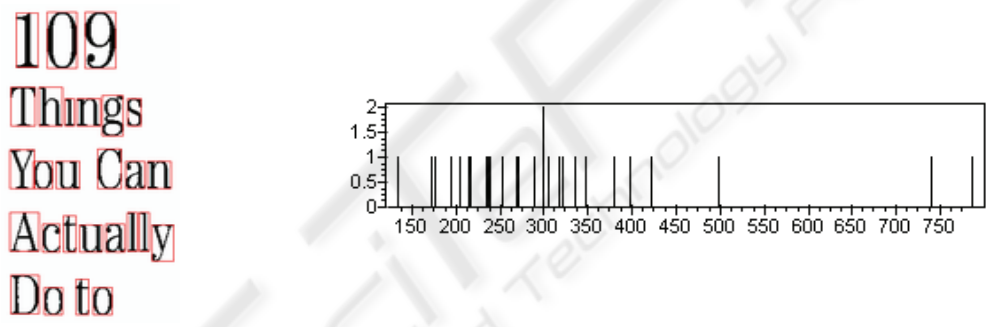
$$HBD(i, j) = \max\{Xl_i, Xl_j\} - \min\{Xr_i, Xr_j\} \quad (3)$$

$$VBD(i, j) = \max\{Yl_i, Yl_j\} - \min\{Yr_i, Yr_j\} \quad (4)$$



(a) (b)

Figure 1: Connecting CCs: (a) Definition of $R(CC_i)$. (b) Experimental example where connections between CCs and $R(CC_i)$ are shown.



(a) (b)

Figure 2: (a) Clean text image with significant character variations, (b) size histogram of (a). The smaller character has $psize = 134$ and the largest $psize = 786$ (5.86 size ratio).

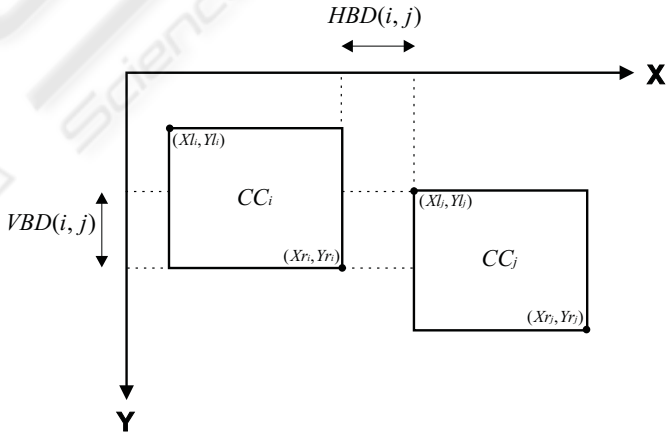


Figure 3: Definition of $HBD(i,j)$ and $VBD(i,j)$ between two CCs.

Figure 4: (a) Original color document, (b) After color reduction (4 colors), (c) Color plane 1, (d) initial grouping, (e) false connections removal, (f) final text localization after superimposition of all color planes.

Equations (3) and (4) are graphically depicted in Fig. 3. Note that if $HBD(i,j) < 0$, CC_i and CC_j overlap in the vertical direction and if $VBD(i,j) < 0$ (as in Fig. 3) they overlap in the horizontal direction.

From this, we state that if

$$HBD(i,j) \geq 0 \quad \wedge \quad VBD(i,j) \geq 0 \tag{5}$$

a connection between CC_i and CC_j is not possible. In words, equation (5) indicates that no overlapping occurs in any of the two directions. We note that connections are bidirectional, which means that connection conditions must apply both for CC_i towards CC_j and the opposite. This can be seen clearly in Fig. 1(b) where character "y" connects with character "n" but not "n" with "y".

The result of the initial grouping process is the creation of connection sets associated to each CC.

$$C(CC_i) = \{c_1^i, ..., c_{cn}^i\} \tag{6}$$

where cn the number of connections for CC_i. If $C(CC_i) = \varnothing$ then no match exists in $R(CC_i)$. Any CC having this property is considered as a non text object (isolated). Thereby, further filtering of non text objects is achieved. An example of the initial grouping procedure is shown in Fig. 4(d). In most cases, characters are assigned with more than four connections. This helps in gathering more information about CCs than taking into account only the closest neighbors.

4.2 Assigning the Direction of Connection Property to CCs

Based on the results of initial grouping, we proceed to the characterization of CCs with a property named Direction of Connection (DOC). The purpose of this strategy is to create homogenous groups of elements, that is to spatially discriminate textual from non textual elements in order for the classification stage (section 5) to be able to be applied and classify text groups from non text groups. Additionally, it supplies the classification module with the information on which it will be based to finally extract text blocks. To define the DOC property, the following two metrics are introduced

$$H_o(CC_i) = -\sum_{j=1}^{cn} VBD(i,j) \tag{7}$$

$$V_o(CC_i) = -\sum_{j=1}^{cn} HBD(i,j) \tag{8}$$

$H_o(CC_i)$ and $V_o(CC_i)$ measure the total amount of CCs overlapping in horizontal and vertical direction, respectively. The DOC property is defined as

$$DOC(CC_i)=\begin{cases}1, & if\ \left(\max\{H_o,V_o\}=H_o\right)\wedge\left(\dfrac{H_o}{V_o}\right)>T_o\wedge\left(H_o\geq H_i\right)\\ 2, & if\ \left(\max\{H_o,V_o\}=V_o\right)\wedge\left(\dfrac{V_o}{H_o}\right)>T_o\wedge\left(V_o\geq W_i\right)\end{cases} \quad (9)$$

$DOC(CC_i)=1$ indicates that CC_i belongs to an horizontal structure and $DOC(CC_i)=2$ that CC_i belongs to a vertical structure. T_o is set at 2 and represents the overlapping difference between horizontal and vertical direction that a CC must have in order to be characterized. It should be noticed that for a tolerance of about 15 degrees (in either horizontal or vertical direction) a significant amount of overlapping between CCs remains and thus these blocks can also be characterized. Depending on these results, the algorithm removes all invalid connections of CCs aiming to the final blocks formation. For example, if CC_i has $DOC(CC_i)=1$ and $VBD(i,j)\geq 0$ then it may not preserve a connection with CC_j (no horizontal overlapping). That is, CC_i belongs to an horizontal structure with no horizontal overlapping with CC_j. The same assumption applies for CCs having $DOC(CC_i)=2$ with $HBD(i,j)\geq 0$. With this technique textual blocks are spatially discriminated from non textual blocks and the final block classification stage can be applied. Fig. 4(e) shows the result of the procedure described in this section. As it can be seen, text elements are grouped in the sense of text lines.

5 CLASSIFICATION OF BLOCKS

In this final step, a classification procedure is applied which classifies the textual and non textual blocks. After the CCs characterization, it is possible some blocks to contain CCs whose DOC is different, that is some have $DOC(CC_i)=1$ and others have $DOC(CC_j)=2$. These are referred as mixed blocks and are considered to be non text. Due to the fact that text block elements are very likely to be assigned with DOC values 1 or 2, a statistical metric is used to reflect this. Let B_j be a structure block containing N CCs, and

$$BH_j=\{CC_i\in B_j : DOC(CC_i)=1,\ i=1,...,k\} \quad (10)$$

represents the subset containing all the CC_i of the structure block B_j that have $DOC(CC_i)=1$. The entire structure block B_j is considered to be an horizontal text block if

$$\frac{k}{N}\geq T_p \quad (11)$$

where $T_p\in[0.5,0.9]$. We apply the same procedure for the vertical structure blocks using the same threshold value.

Applying the above classification procedure for the image of Fig. 4, the final text localization results after superimposition of all color planes are shown in Fig. 4(f).

6 EXPERIMENTAL RESULTS

For the present work, a large image database of color documents was created (1000 images at 150-300 dpi). Some images were scanned from color book covers and journals and some were obtained from the internet. The proposed technique was extensively tested with different types of color documents.

To measure the performance of our algorithm, text block precision rate (BPR) was used.

$$BPR=\frac{N_c}{N_e} \quad (12)$$

where N_c is the correct extracted text blocks and N_e the total number of extracted blocks. We have found a mean value of $BPR\approx 85\%$.

In Fig. 5 we present an experimental result of text localization of a color document. Fig. 5(b) shows the result of the color quantization procedure which has produced an image with 7 colors. To demonstrate the element grouping stage we show its application on a color plane that contains text and non text CCs. Fig 5(c) depicts the result of the initial grouping stage that has been described in section 4.1. Note that some formed groups contain text and non text elements. Also for some CCs no match was found so these were removed from further examination. The final grouping stage as described

in section 4.2 is shown in Fig 5(d), where the groups of the initial grouping procedure are refined in order to form the final groups which will be classified with the use of the information obtained by the DOC of each CC. Final extracted areas, from all color planes, considered as text are shown in Fig 5(e).

The second experimental result is presented in Fig. 6 which is a color document of a book cover. In this example vertically and horizontally aligned text coexists. The grouping of the CCs that corresponds to the color plane of the vertically aligned text is shown in Fig. 6(c) and Fig. 6(d). The final result (Fig. 6(e)) shows that the method successfully extracted text of both orientations.

7 CONCLUSIONS

Text localization is an important processing in computer vision systems, especially for document or text image related applications. In this paper, we have presented a new method for text localization which is suitable for complex cover pages and any type of color documents. In these cases of document images text and graphics are highly mixed with the background. The proposed technique efficiently integrates a KSOM color quantization procedure and a color plane text localization technique. The proposed technique is robust and has the following characteristics:

- Preferable estimation of the number of dominant colors
- Color reduction by using the KSOM neural network.
- Splitting the color document image into a number of binary images, called color planes, corresponding to the dominant colors obtained.
- In every color plane, CCs are spatially formed in groups with the use of local information in an adaptively defined area.
- The information that is used of each CC for classification involves not only the closest neighbors but a large number of similar CCs.

First results on a large number of data set of above 500 color text images collected from Internet, most of which are color cover pages, are very encouraging.

REFERENCES

Atsalakis, A., Papamarkos, A. and Andreadis, I., 2002. On estimation of the number of image principal colors and color reduction through self-organized neural networks, Int. Journal of Imaging Systems and Technology, 12(3), 117-127.

Chen, W.Y. and Chen, S.Y., 1998. Adaptive page segmentation for color technical journals' cover images, Image and Vision Computing, 16(12-13), 855-877.

Fletcher, L. and Kasturi, R., 1988. A robust algorithm for text string separation from mixed text/graphics images, IEEE Trans. PAMI, 10(6), 910–918.

Hase, H., Shinokawa, T., Yoneda, M. and Suen, C.Y., 2001. Character string extraction from color documents, Pattern Recognition, 34(7), 1349–1365.

Jain, A.K. and Zhong, Y., 1996. Page Segmentation Using Texture Analysis, Pattern Recognition, 29(5), 743-770.

Jain, A.K. and Bhattacharjee, S., 1992. Text segmentation using Gabor Filters for automatic document processing, Mach. Vision Appl., 5, 169–184.

Jung, K. and Han, J., 2004. Hybrid approach to efficient text extraction in complex color images. Pattern Recognition Letters, 25(6), 679-699.

Jung, K., Kim, K.I. and Jain, A.K., 2004. Text information extraction in images and video: A survey, Pattern Recognition, 37(5), 977-997.

O'Gorman, L., 1993. The Document Spectrum for Page Layout Analysis, IEEE Trans. PAMI, 15(11), 1162-1173.

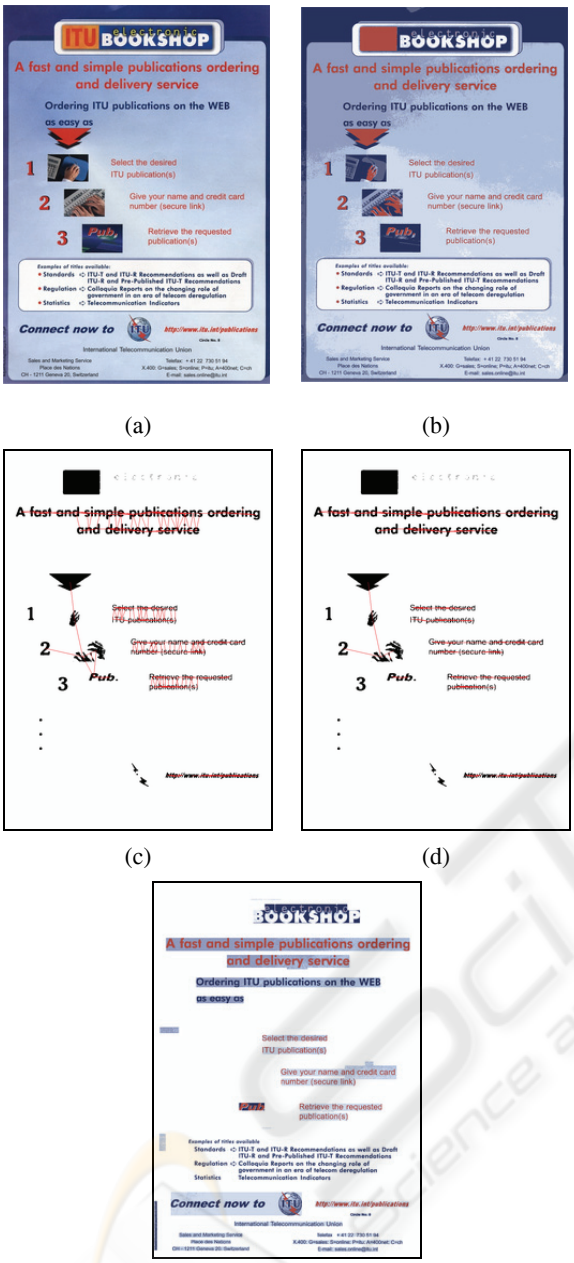
Papamarkos, N., 1999. Color reduction using local features and a SOFM neural network, Int. Journal of Imaging Systems and Technology, 10(5), 404-409.

Simon, A., Pret, J.C. and Johnson, A.P., 1997. A Fast Algorithm for Bottom-Up Layout Analysis, IEEE Trans. Pattern Analysis and Machine Intelligence, 19(3), 273-277.

Sobottka, K. et al., 2000. Text Extraction from Colored Book and Journal Covers, International Journal on Document Analysis and Recognition, 2(4), 163-176.

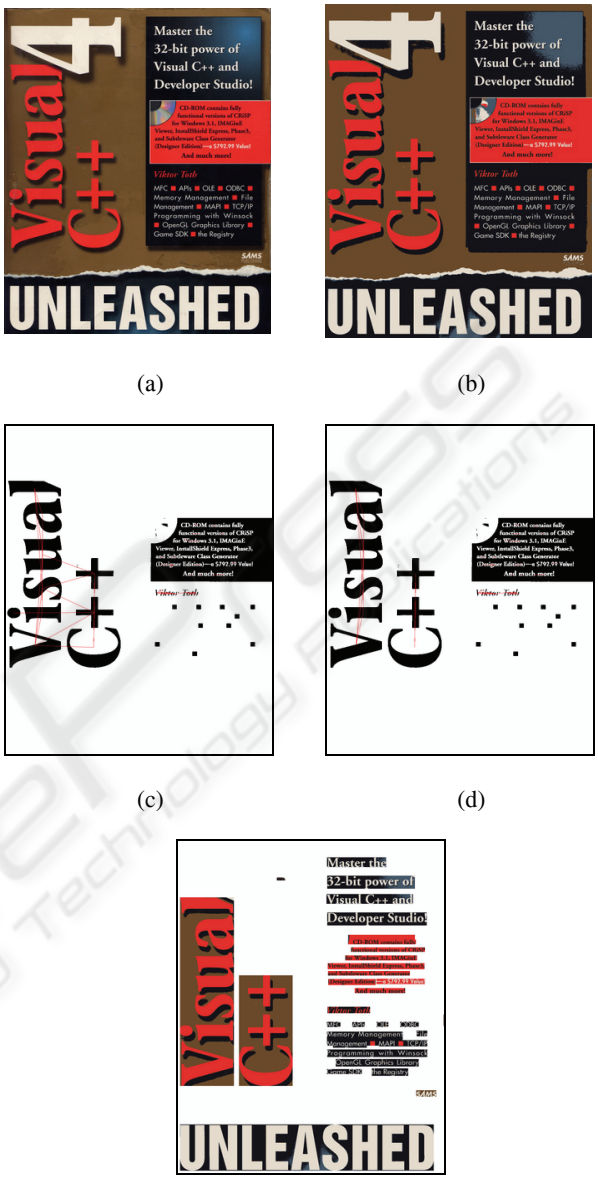
Strouthopoulos, C., Papamarkos, N. and Atsalakis, A., 2002. Text extraction in complex color documents. Pattern Recognition, 35(8), 1743-1758.

Zhong, Y., Karu, K., Jain, A.K., 1995. Locating text in complex color images, Pattern Recognition, 28 (10), 1523–1535.



(a) (b) (c) (d) (e)

Figure 5: Text localization example. (a) original image (b) original image after color reduction (7 colors) (c) initial elements grouping for color plane 7 (d) final elements grouping (e) text localization result after classification of groups and superimposition of all color planes ($T_p = 0.55$).



(a) (b) (c) (d) (e)

Figure 6: Text localization example. (a) original image (b) original image after color reduction (5 colors) (c) initial elements grouping for color plane 4 (d) final elements grouping (e) text localization result after classification of groups and superimposition of all color planes ($T_p = 0.8$).