Text Extraction Using Component Analysis and Neuro-fuzzy Classification on Complex Backgrounds

Michael Makridis[1], Nikolaos E. Mitrakis[2],
Nikolaos Nikolaou[1], and Nikolaos Papamarkos[1]

[1] Image Processing and Multimedia Laboratory, Department of Electrical & Computer Engineering, Democritus University of Thrace, 67100 Xanthi, Greece
[2] European Commission, Joint Research Centre, Institute for Protection and Security of the Citizen, Maritime Affairs Unit G.04, TP 051, 21027 Ispra (VA), Italy
{mmakridi,papamark,nnikol}@ee.duth.gr,
nikolaos.mitrakis@jrc.ec.europa.eu

Abstract. This paper proposes a new technique for text extraction on complex color documents and cover books. The novelty of the proposed technique is that contrary to many existing techniques, it has been designed to deal successfully with documents having complex background, character size variations and different fonts. The number of colors of each document image is reduced automatically into a relative small number (usually below ten colors) and each document is divided into binary images. Then, connected component analysis is performed and homogenous groups of connected components (CCs) are created. A set of features is extracted for each group of CCs. Finally each group is classified into text or non-text classes using a neuro-fuzzy classifier. The proposed technique can be summarized into four consequent stages. In the first stage, a pre-processing algorithm filters noisy CCs. Afterwards, CC grouping is performed. Then, a set of nine local and global features is extracted for each group and finally a classification procedure detects document's text regions. Experimental results prove the efficiency of the proposed technique, which can be further extended to deal with even more complex text extraction problems.

Keywords: Text extraction, Color reduction, Connected component analysis, Adaptive run length smoothing, Pattern classification, Neuro-fuzzy classifier.

1 Introduction

This paper proposes a technique for text extraction in complex color documents and cover books. Interest about exploiting text information in images and video has grown notably during the past years. Text can provide powerful description of the image content and it reasonably attracts the research interest.

A main categorization of text extraction methods include texture based techniques [1]-[4] and connected components (CCs) based techniques [5]-[9].

Texture based methods use the observation that text in images has distinct textural properties that distinguish them from the background. They are mainly used in video text based applications [10]-[12]. On the other hand, CCs based techniques are fast

A. Heyden and F. Kahl (Eds.): SCIA 2011, LNCS 6688, pp. 742–751, 2011.

and relatively simple in implementation and exploit the fact that characters are segmented. The proposed approach belongs to this specific category of text information extraction techniques.

The proposed technique performs color reduction to limit document's colors and divides each document into a set of binary documents, one for every color. Then, it performs connected component (CC) analysis and creates groups of components. For each group, a set of features is extracted and finally the classification process, based on a neuro-fuzzy system, detects those groups that correspond to text regions.

Most text extraction techniques focus on data sets of documents with certain specifications such as:

- Documents pixel depth is 8-bit gray-scale
- Documents have low character size variations
- Text gray values are greater than background values
- Documents have uniform background without contrast variations

The novelty of the proposed paper is that overcomes the specifications mentioned above and deals successfully with complex text extraction problems.

2 Description of the Technique

The technique proposed in this paper is based on an iterative procedure of four stages. A document image is the input of the technique. The number of its colors is decreased (usually in less than ten colors) according to a color reduction technique [7]. After color reduction, the initial document can be represented by a set of binary images, one for each color, which we call color planes. Then, an iterative procedure is applied to each color plane. Generally, the proposed technique can be summarized into four stages:

Stage 1. Pre-processing: CC analysis is performed to each color plane. Color reduction process usually creates noisy, superfluous CCs. Most of these CCs, though, can be easily recognized and removed during this stage.
Stage 2. Page segmentation: CCs of each color plane are grouped according to an adaptive run length smoothing algorithm (ARLSA) [16].
Stage 3. Feature extraction: Each group of CCs is considered as a pattern. For each pattern, a set of nine local run length and spatial features is extracted.
Stage 4. Classification using Adaptive Neuro-Fuzzy Interference System (ANFIS) [14]. A subset of patterns is first used to train the classifier.

A block diagram of the proposed technique is shown in Fig. 1. In the rest of this section a brief description of color reduction [7] and ARLSA [13] algorithms is given.

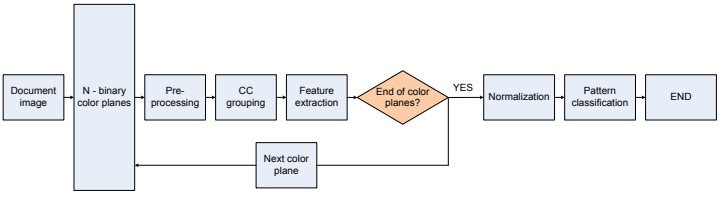

Fig. 1. The block diagram of the proposed technique

2.1 Color Reduction

A color document or a cover book has millions of different color values. In order to apply CC analysis, we have to limit the total number of colors. To achieve that, we use an unsupervised clustering algorithm to find clusters of similar colors, originally proposed by Sobottka et al. [7]. We chose to implement this color reduction technique for three basic reasons:

- Simplicity of the algorithm
- Very low computational cost
- Text objects (CCs) are coherent and final color distribution inside the document image is homogenous.

2.2 Adaptive Run Length Smoothing Algorithm

Adaptive run length smoothing algorithm (ARLSA) [13] is a modified version of RLSA [15], a common algorithm, that it is used in page layout analysis and segmentation techniques. Generally, ARLSA is applied on CCs of binary images.

$$\begin{aligned} L(CC_i, CC_j) &< T_l \\ H_R(CC_i, CC_j) &< T_h \\ O_R(CC_i, CC_j) &< T_o \end{aligned} \tag{1}$$

The novelty of ARLSA is, that it applies run length at a certain direction only between pixels of different CCs and only if these CCs fulfill certain specifications. In the proposed technique, ARLSA is applied in the horizontal direction.

Let p_i and p_j be two pixels that belong to CCs CC_i and CC_j and $CC_j\,(i \neq j)$. The connection between p_i and p_j is made only if the following specifications are fulfilled:

where $L(CC_i, CC_j)$ is the Euclidean distance between the bounding boxes of CC_i and CC_j, $H_R(CC_i, CC_j)$ is the height ratio between CC_i and CC_j and $O_R(CC_i, CC_j)$ is the overlapping ratio between CC_i and CC_j. Furthermore, H_R can be defined as:

$$H_R(CC_i, CC_j) = \min(H_{CC_i}, H_{CC_j}) / \max(H_{CC_i}, H_{CC_j}) \tag{2}$$

Where H_{CC_i} and H_{CC_j} the heights of CC_i and CC_j.

Finally, O_R is defined as the overlapping ratio between two components.

(a) (b) (c)

Fig. 2. ARLSA filtering example: (a) A color plane, (b) color plane after application of ARLSA, (c) filtered color plane

3 Image Pre-processing

The purpose of this stage is to remove small noisy connected components and large background or graphic components. Pre-processing filtering is applied in two separate steps:

First, noisy elements are filtered out based on three characteristics of the connected components and their corresponding bounding boxes. For a connected component CC_i these characteristics are:

The height of the bounding box of the CC_i, H_{CC_i}

The elongation $E(CC_i) = \dfrac{\min\{H_{CC_i}, W_{CC_i}\}}{\max\{H_{CC_i}, W_{CC_i}\}}$

The density $D(CC_i) = \dfrac{P_{num}(CC_i)}{BB_{size}(CC_i)}$,

which is the ratio of the number of foreground pixels $P_{num}(CC_i)$ to the total number of pixels in the bounding box $BB_{size}(CC_i) = H(CC_i) \cdot W(CC_i)$.

Connected components with $H(CC_i) < AH/3$, or $D(CC_i) < 0.08$, or $E(CC_i) < 0.08$ are considered as noisy elements and they are eliminated, where AH is the average height of all CC of the color plane. These values have been selected very carefully, so no character elements will be eliminated.

The second type of filtering removes large background and graphic components. It is based on the comparison of the connected components from two images, the initial binary document image and the resulted image after the application of the ARLSA. Let I_1 be the original image (see Fig. 2(a)), I_2 the image after the application of the ARLSA (see Fig. 2(b)). The number of pixels P_{I_2} of each connected component $CC_i \in I_2$ is calculated, that is the number of the black pixels. In the defined area of each $CC_i \in I_2$, the sum P_{I_1} of the corresponding black pixels of I_1 is also calculated.

The ratio of these two sums is taken into account as in the following equation:

$$P_R = \frac{P_{I_2}}{P_{I_1}} \leq T_R \tag{3}$$

As it is mentioned above, ARLSA connects only similar neighbor components. In this case, graphic components of Fig. 2(a) are isolated and the application of ARLSA does not link them with other components. Therefore, their pixel size remains almost the same and P_R has a value near to one. Components, which correspond to pixel size ratio smaller than T_R (Eq. 3), are removed and a new image I_3 (see Fig. 2(c)) is produced. The parameters used in ARLSA are those parameters proposed as optimal by the authors.

4 Document Segmentation

Document page segmentation is very important for successful classification. During this stage, CCs of color planes are grouped to form a pattern for the classification procedure. False grouping will have as a result unreliable feature values and furthermore classification failure. Therefore, we need a reliable technique that groups CCs of the same class (text or non-text).

To perform successful grouping, we use ARLSA. ARLSA groups only similar CCs as far as height and overlapping is concerned.

The choice of ARLSA is based on the following reasons:

- Characters are in most cases CCs of similar height in a certain direction (in most cases horizontal). Furthermore, the height ratio between two characters of the same font size (in the same sentence) is less than 2.
- Graphics consist of CCs that have great variation in height. Furthermore, graphic CCs do not have a defined arrangement in space and therefore overlapping measure in a certain direction is very low.
- Background CCs are large isolated CCs.

Because of the above reasons, text CCs group together in most cases, while non-text CCs form small groups. Each group is considered as a pattern for the feature extraction and classification stages.

5 Feature Extraction

In this stage, we form a set of nine features for each pattern (group of CCs) of each binary color plane. Feature selection has been made carefully, in order to distinguish text from non-text patterns as much as possible.

Mean Elongation: Elongation feature has been introduced in Section 3. Each pattern after CC grouping is formed by a set of CCs. The value of this feature is the mean elongation of a pattern's CCs. The idea of choosing mean elongation is that usually, character CCs have similar width to height ratio. On the other hand, lines and big graphic CCs can have either too small or too big width to height ratio.

Mean Density: Density feature has been also introduced in Section 3. The value of this feature is the mean density of a pattern's CCs. The idea of choosing mean density is that most graphic CCs have many holes and therefore their density values are smaller than character CCs.

Mean pixel size: This feature represents the mean pixel size of the CCs of each pattern.

Local Connectivity: This feature measures the coherence of a pattern. For each pixel $p_{i,j}$ of a pattern, the number of neighbor pattern pixels, within a 3x3 neighborhood, is counted. The total number of neighbor pattern pixels is divided by the pixel size of the pattern for normalization reasons. This feature takes large values for coherent CCs, while it takes small values for thin CCs or CCs with many holes. It can be expressed as follows:

$$lc_l = \frac{\sum_{k=1}^{PS_l} \sum_{m=i-1}^{i+1} \sum_{n=j-1}^{j+1} p_{l,i,j}}{PS_l} \tag{4}$$

Where $p_{l,i,j}$ a pixel of a pattern l and PS_l the total number of pixels of the pattern, which is pixel size of pattern l.

Run length mean and Run length variance features: For each pattern, we calculate the mean run length value at a certain direction that we call it G_d (group direction). Each pattern is a group of CCs, as it is mentioned in Section 4. The center points of these CCs define a least squares line. The gradient of this line, we call it G_d. Least squares line is represented by the following equation:

$$y = a + bx \tag{5}$$

We are interested in the direction of this line, which is defined as:

$$b = \frac{n\sum_{l=1}^{n} Xc_l Yc_l - (\sum_{l=1}^{n} Xc_l)(\sum_{l=1}^{n} Yc_l)}{n\sum_{l=1}^{n} Xc_l^2 - (\sum_{l=1}^{n} Xc_l)^2} \tag{6}$$

Where n is the number of a pattern's CCs and $\{(Xc_0 Yc_0),...,(Xc_n Yc_n)\}$ are their center points.

Run length mean and variance features are extracted from the run length histogram of each pattern.

Group direction mean feature: For each pair of CCs of a pattern . CC_i . and CC_j, we compute the corresponding gradient $G_{d,i,j}$. Group direction mean feature is the mean value of all $G_{d,i,j}$.

Mean Overlapping feature: As it is mentioned above, the center points of the CCs of a pattern define a group direction G_d. For each pair of CCs of a pattern CC_i and CC_j, we compute the corresponding overlap measure in the direction of G_d. The mean overlapping feature is the mean overlapping value of a pattern.

Fig. 3 illustrates an example of this feature. Suppose that the word "Example" is a pattern that consists of seven CCs, we calculate the overlap between letters-CCs "E" and "X" in the group direction. This feature is similar to the overlap feature that we introduced in Section 2.2 in the horizontal direction, but now overlapping is calculated in the direction of G_d

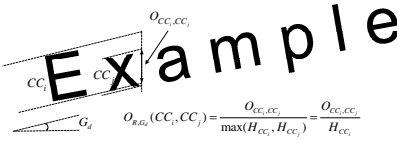

Fig. 3. Overlapping feature for CCs "E" and "x" in the direction of G_d

Black and white alterations: This feature is calculated by the total number of alterations between pattern and non-pattern pixels in the group direction, G_d .The number of black and white alteration is divided by the width of the pattern for normalization.

6 Pattern Classification Using ANFIS

ANFIS (Adaptive-Network-based Fuzzy Interference System) [14] is a neuro-fuzzy multilayered architecture, which was first introduced by Jang, well known for dealing with complex nonlinear modeling or classification problems. ANFIS main advantage is that it combines the strong descriptive characteristics of fuzzy logic with the learning capabilities of neural networks.

ANFIS consists of 6 layers which are described below:

Layer 1: The nodes of this layer carry the inputs of the network to the next layer.

Layer 2: Each node of this layer implements a fuzzy membership function that describes a fuzzy set of each input (linguistic nodes). The proposed implementation uses Gaussian membership functions which are described by the following equation:

$$\mu_{A_j^i}(x_j) = e^{\frac{-(x_j-\sigma_j^i)^2}{2\sigma_j^{i2}}}, \quad j = 1,\ldots,m, \qquad i = 1,\ldots,k_j \tag{7}$$

The output of this layer reveals the membership degree of feature x_j to fuzzy set A_j^i, where j stands for the input and i for the fuzzy set defined in input j. The fuzzy input partition has been implemented through subtractive clustering [16].

Layer 3: The nodes of this layer are called rule nodes. The output of each node represents the degree that satisfies the hypothesis of a rule. The number of nodes of this layer is equal to the number of the rules, that is $n = k_1 \times \cdots \times k_m$. The degree of fulfillment of each rule can be calculated by the following:

$$\mu_i(\boldsymbol{x}) = \prod_{j=1}^{m} \mu_{A_j^i}(x_j) \text{ , i = 1,...n} \tag{8}$$

Layer 4: In this layer, the normalized fulfillment of each rule is calculated:

$$\overline{\mu}_i(\boldsymbol{x}) = \frac{\mu_i(\boldsymbol{x})}{\sum_{i=1}^{n} \mu_i(\boldsymbol{x})} \text{ , i = 1,...n} \tag{9}$$

Layer 5: The nodes of this layer calculate the output of each rule:

$$y(i) = \overline{\mu}_i(\boldsymbol{x}) \times w_i \text{ , i = 1,...n} \tag{10}$$

Layer 6: In this layer the node calculates the final output of the model by summarizing the partial output of each rule:

$$y = \sum_{i=1}^{n} y(i) \tag{11}$$

Fuzzy input partition has been implemented via subtractive clustering [16], while hybrid batch learning algorithm [15] is used to calculate the parameters of the network.

7 Experimental Results

In order to achieve objective experimental results, we created a dataset with document and ground truth images for evaluation purpose. Document dataset consists of 50 color cover books and documents with complex background, various font colors, sizes and types. Ground truth binary images were created for all 50 documents manually, using commercial image processing software. The proposed technique can identify text areas with skew up to 45 degrees. The documents are all taken from the internet, while their resolution is at least 200 dpi.

Due to space limitations, we present in Fig. 4 two characteristic results that they should be discussed. These examples reveal some of the advantages and disadvantages of the proposed technique.

Fig. 4 (b) shows a successful result. The cover book has non-uniform background and fonts of different sizes and colors. Main contribution to the successful result has the great resolution and successful color reduction that leads to coherent CCs.

Fig. 4 (c) shows a movie poster with uniform background and a large graph. Text areas, which include fonts of different color, type and size, have been successfully detected. However, some graph patterns in the middle of the Fig. 4 (d) have wrongly classified as text. These patterns have common text characteristics, such as elongation, density, run length variation and overlapping, which lead to classification error.

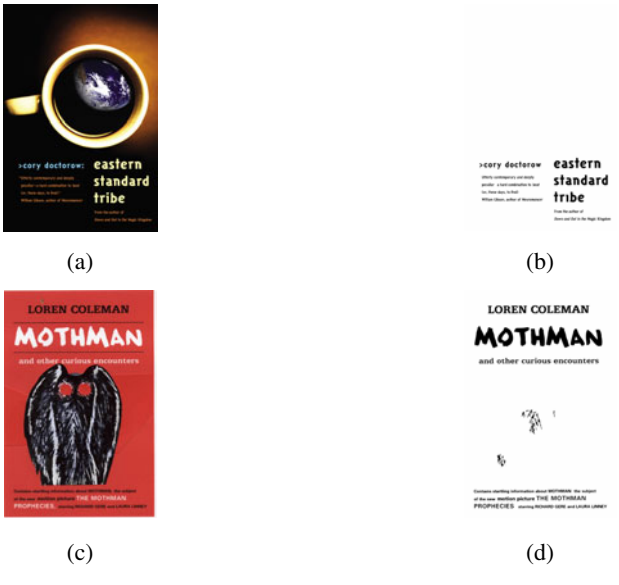

Fig. 4. Text extraction examples: (a), (c) Original document images, (b), (d), resulted document images

8 Conclusions

We have presented a new technique for text extraction on complex color documents. In this type of documents, text and graphics are highly mixed with the background and therefore color reduction, page segmentation and furthermore text extraction is a challenging task. Experiments have been performed and presented to test the effectiveness of the proposed technique.

The main advantage of the presented technique lies on the fact that although it deals with complex documents, it performs high successful rates and a reliable result for further processing. Pattern extraction based on connected component analysis and classification using ANFIS seem to work fine, even under extreme circumstances. Additionally, ARLSA provides very useful information about text patters. However, we intend to perform more research in the field of color reduction (extraction of color planes) and pattern extraction.

References

1. Jain, A.K., Zhong, Y.: Page Segmentation Using Texture Analysis. Pattern Recognition 29, 743–770 (1996)
2. Wu, V., Manmatha, R.: TextFinder: an automatic system to detect and recognize text in images. IEEE Trans. on Pattern Analysis and Machine Intelligence 21, 1224–1229 (1999)
3. Deng, S., Latifi, S., Regentova, S.: Document segmentation using polynomial spline wavelets. Pattern Recognition 34, 2533–2545 (2001)
4. Wang, B., Li, X., Liu, F., Hu, F.: Color text image binarization based on binary texture analysis. Pattern Recognition Letters 26, 1650–1657 (2005)
5. Fletcher, L., Kasturi, R.: A robust algorithm for text string separation from mixed text/graphics images. IEEE Trans. on Pattern Analysis and Machine Intelligence 10, 910–918 (1988)
6. Chen, W.Y., Chen, S.Y.: Adaptive page segmentation for color technical journals' cover images. Image and Vision Computing 16, 855–877 (1998)
7. Sobottka, K., Kronenberg, H., Perroud, T., Bunke, H.: Text Extraction from Colored Book and Journal Covers. International Journal on Document Analysis and Recognition 2, 163–176 (2000)
8. Hase, H., Shinokawa, T., Yoneda, M., Suen, C.Y.: Character string extraction from color documents. Pattern Recognition 34, 1349–1365 (2001)
9. Strouthopoulos, C., Papamarkos, N., Atsalakis, A.: Text extraction in complex color documents. Pattern Recognition 35, 1743–1758 (2002)
10. Lyu, M.R., Song, J., Cai, M.: A comprehensive method for multilingual video text detection, localization, and extraction. IEEE Trans. on Circuits and Systems for Video Technology 15, 243–255 (2005)
11. Chen, Y.L., Wu, B.F.: Text extraction from complex document images using the multiplane segmentation technique. In: Proceedings of IEEE International Conference on Systems, Man and Cybernetics, vol. 4, pp. 3540–3547 (2006)
12. Xu, L., Wang, K.: Extracting text information for content-based video retrieval. LNCS, pp. 58–69 (2008)

13. Nikolaou, N., Makridis, M., Gatos, B., Stamatopoulos, N., Papamarkos, N.: Segmentation of Historical Machine-Printed Documents Using Adaptive Run Length Smoothing and Skeleton Segmentation Paths. Image and Vision Computing (2009)
14. Jang, J.-S.R.: ANFIS: adaptive-network-based fuzzy inference system. IEEE Transactions on Systems, Man, and Cybernetics 23, 665–685 (1993)
15. Wahl, F.M., Wong, K.Y., Casey, R.G.: Block Segmentation and Text Extraction in Mixed Text/Image Documents. Computer Graphics and Image Processing 20, 375–390 (1982)
16. Chiu, S.: Fuzzy Model Identification Based on Cluster Estimation. Journal of Intelligent & Fuzzy Systems 2, 267–278 (1994)