Text Extraction in Complex Color Documents

Pattern recognition, Vol. 35, No. 8, August 2002
C. Strouthopoulos, N. Papamarkos, A. E. Atsalakis

School of Electrical Engineering and Computer Science
Kyungpook National Univ.
Text extraction in complex color documents

Adaptive Color Reduction

Page Layout Analysis

Merging the Text Regions

Final document

Mixed-type color document
Abstract

◆ A new method to automatically detect and extract text in mixed-type color documents
  – Combination of an adaptive color reduction (ACR) and a page layout analysis (PLA)
  – ACR
    • To obtain the optimal number of colors
    • To convert the document into the principal of them
  – PLA
    • Applying independently to each of the color plains
    • Identifying the text region
  – Merging procedure
1. Introduction

◆ Page layout analysis
  – Pre-processing stage
    • Optical character recognition
    • Document retrieval
    • Document compression
  – Key phase
    • Segmentation of a mixed-type document into text and non-text regions
  – Difficulties associated with the large number of unique colors   Color quantization algorithm
New method for text extraction

1\textsuperscript{st} stage
- Unsupervised adaptive color reduction (ACR)
  - Unsupervised neural network classifier
  - Tree-search procedure
- Split and merging condition

2\textsuperscript{nd} stage
- Color planes $\rightarrow$ binary document
- PLA technique
  - Run length segmentation algorithm (RLSA)
  - Neural network block classifier (NNBC)
- Merging procedure
- 3rd stage
  - Self-organized feature map (SOFM)
    - Colors of the text regions reduced to two
  - The characters appear with solid colors in the segmented document
2. Optimal color reduction

◆ Color reduction method – ACR
  – Color reduction using an adaptive tree clustering procedure
  – In each node of tree, a self-organized neural network classifier (NNC) is used
    • Image color values
    • Additional local spatial features
  – NNC = PCA + Kohonen SOFM neural network
    • PCA : Principal component analyzer
    • SOFM : Self-organized feature map
The ACR technique

- A color space = 3D vector space
- General color image function

\[
I_k(i, j) = \begin{cases} 
  c_1(i, j) & \text{if } k = 1 \\
  c_2(i, j) & \text{if } k = 2 \\
  c_3(i, j) & \text{if } k = 3 
\end{cases}
\]  

(1)

- \(N(i, j)\): Local neighboring region of pixel \((i, j)\)
- \(f_k\): Values of the colors of \(N(i, j), k = 4, ..., K + 3\)
– A tree scheme of the ACR

Fig. 1. A tree scheme of the adaptive color reduction algorithm.
– The feature extraction procedure
- The classifier used in each node

![Diagram of neural network classifier]

**Fig. 2.** The structure of the neural network classifier.
Split and merging conditions

- Applying locally to each tree node during the adaptive clustering process

- Split conditions
  - The variance of each class; less than a threshold value
  - The variance of the class centers; less than a threshold value
  - The increasing of the variance
  - The minimum number of pixels in each class
  - The minimum number of samples; exceed an upper bound
◆ Image sub-sampling
  – In the case of large-size documents
  – In order to achieve reduction of the computational time and memory size requirements
  – Fractal scanning process based on the Hilbert space filling curve

\[
\frac{4^k}{\text{number of image pixels}}
\]
The stages of the method

- Define the desired max. number J
- Define the type of the spatial features
- Define the split and merging conditions
- Construct the PCA and SOFM neural network
- Define the sub-sampling parameters
- In each tree node, train the neural network, transform the colors of the original image
- Construct the final image by merging the separated color regions obtained
3. Identification of text areas

- Modified version of PLA
  - Block extraction
    - Applying RLSA globally and locally to each color plane
  - Block classification
    - NNBC: text, drawing or graphic blocks
  - Additional classification criteria
    - Rejecting blocks constructed of noise blocks
    - Classifying the white character blocks as non-text blocks
Fig. 3. The original color document.
Table 1.
RGB values of the four principle colors

<table>
<thead>
<tr>
<th>R</th>
<th>G</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>91</td>
<td>88</td>
<td>87</td>
</tr>
<tr>
<td>194</td>
<td>183</td>
<td>176</td>
</tr>
<tr>
<td>193</td>
<td>227</td>
<td>243</td>
</tr>
<tr>
<td>248</td>
<td>247</td>
<td>250</td>
</tr>
</tbody>
</table>

Fig. 4. Binary documents corresponding to the four color planes.
Block characteristics

• $A = WH$: Block area = block width x block height
• The number $O_p$ of the object pixels
• The number $C$ of transmissions
  - “0” to “1” and “1” to “0” (1: object / 0: background)
• The number $V_L$ of top to bottom vertical complete background lines

Non-text block if at least one of the criteria is satisfied

\[ \frac{V_L}{W} < T \quad T = 0.01 \quad (2) \]

\[ \frac{O_p}{A} < 0.07 \quad \text{or} \quad \frac{O_p}{A} > 0.3 \quad (3) \]
\[ \frac{C}{O_p} < 0.25 \quad \text{or} \quad \frac{C}{O_p} > 2 \] 

(4)

- PLA technique
  - Separation of the document into 8x8, identification of text regions using NNBC
  - Global application of the RLSA
  - Identification of high marks
  - Classification of high marks by using NNBC
  - Local application of RLSA
  - Filtering the extracted blocks
  - Classification of the rest blocks as text or non-text regions
Fig. 5. (a) Text regions identified in the binary document of Fig. 4(a).
(b) Text regions identified in the binary document of Fig. 4(c).
4. Merging text areas

- Improving character shapes
  - Text and background colors in each text region are re-defined locally
  - Two principal colors are obtained using a neural network SOFM
  - Obtaining the color corresponds to the character colors
    • ACR technique $\rightarrow b$ one of color planes, $C_b(R_b, G_b, B_b)$
    • $t_b$ a text region in
    • Color classes obtained by the neural network SOFM classifier, $C_1(R_1, G_1, B_1), C_2(R_2, G_2, B_2)$
– Among $C_1$ and $C_2$ the class corresponding to the character color is the class having RGB color nearest to the $C_b$ class

– A pixel $p$ of a text region $t_b$ is considered as a character pixel

Table 2. RGB values of the $C_1, C_2, C_b$ color classes

<table>
<thead>
<tr>
<th>$C_b$</th>
<th>$C_1$</th>
<th>$C_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>R</td>
<td>G</td>
<td>B</td>
</tr>
<tr>
<td>91</td>
<td>88</td>
<td>87</td>
</tr>
<tr>
<td>91</td>
<td>88</td>
<td>87</td>
</tr>
<tr>
<td>91</td>
<td>88</td>
<td>87</td>
</tr>
<tr>
<td>91</td>
<td>88</td>
<td>87</td>
</tr>
<tr>
<td>91</td>
<td>88</td>
<td>87</td>
</tr>
</tbody>
</table>
– Text regions before and after the local estimation

Fig. 6. Text regions before and after the local estimation of the character colors.
INTERLOCUTOR TO CANDIDATES A & B: Here are some pictures of things we use every day. Talk to each other about how important each one is in your life, and then choose three that are so important that you could not live without them.

CANDIDATE A

Well, to be perfectly honest I can’t say a cooker is that important in my life. My mother does all the cooking. A phone, on the other hand, is essential for me. I use it every day to talk to my friends.

Well, maybe you’re right, but years ago people lived without them, didn’t they? I disagree with you about cars, though. I think they are a necessity in modern life. What about a computer? I haven’t got one but most of my friends have, and now wouldn’t be without one.

Yes, that’s true. How about food? That’s probably the most important thing of all.

I suppose you’re right. Anyway, I think a car, a TV and a phone are three things it would be difficult for me to live without.

CANDIDATE B

I see your point. A phone is useful, but as far as I’m concerned not necessary. As for transport, cars are useful, but not essential, in my opinion. I imagine it would be difficult to live without a fridge, though, especially in our climate.

Yes, I know what you mean, but I don’t need one, personally. One thing I could not live without is a TV. It keeps me entertained and informed. A camera, on the other hand, is not exactly essential, unless you’re a photographer, of course.

In a way yes, but that’s junk food and there are better things to eat.

I wouldn’t agree with you about the car, but I definitely could not do without a television and a cooker. My third choice would be a fridge because we don’t have time to go out and buy fresh food every day.
5. Examples

◆ Example 1

Fig. 8. (a) The original document. (b) Document with only five colors.
– Results of example 1

**NEC Direction SM-400B**
- Intel® Pentium® II Processor 400MHz
- 32MB Internal Cache
- 32MB SDRAM
- 448MB Ultra DMA/133 Hard Drive
- 8X DVD-ROM Drive
- Realtek Sound Solution 2200 ACP Audio/AGP Graphics Accelerator with 8MB SCRAM Video Memory
- 17” NEC Colour Monitor 90.85, 90.85, 16.0” viewable
- SoundBlaster Audio PCI 1645 Sound Card
- Speakers: 1,998 Speakers, Headphones 10 Mts.
- Diamond Stealth Modems
- English/French/Italian Keyboard
- Microsoft® Windows 95, 98, Windows Me, Windows 2000, Windows NT
- Norton Utilities
- £899

**NEC Direction SM-500B3**
- Intel® Pentium® III Processor 500MHz
- 32MB Internal Cache
- 64MB SDRAM
- 544MB Ultra DMA/33 Hard Drive
- 8X DVD-ROM Drive
- ATI 3D Radeon 8MB AGP Graphics Accelerator with 32MB DDRII Video Memory
- 19” NEC Monitor 90.85, 90.85, 16.0” viewable
- SoundBlaster Live 5.1 Sound Card
- Speakers: 1,998 Speakers, Headphones 10 Mts.
- Diamond Stealth Modems
- English/French/Italian Keyboard
- Microsoft® Windows 95, 98, Windows Me, Windows 2000, Windows NT
- Norton Utilities
- £1499

---

**Fig. 9. Final text extraction results.**
Fig. 10. (a) The original document. (b) Document with only five colors.
– Results of example 2

**Fig. 11.** Final text extraction results.
Example 3

Fig. 12. (a) Original color document. (b) Document after color quantization.
- ACR results of example 3

Fig. 13. The four color planes of the document.
– Final results of example 3

VerticalNet. In the race for answers, you're either a step ahead...or falling behind.

As a design engineer, you need answers now...not tomorrow. And you don't have time to
dig in a dozen different places just to come up dry, either. That's why VerticalNet was born.
Our Communications and Electronics Group are the online information source dedicated solely
to design engineers, and we're going to revolutionize the way you work...forever.

You can't get more up-to-the-minute information about your industry anywhere; our editorial
content is constantly updated seven days a week. Do you know any magazines, or their website, who can say that?
You can also log on anytime for a complete directory of the leading suppliers in your industry, download the latest software
demo, talk with fellow engineers in our chat rooms, catch up on the hottest product reviews and trends...and more.
Plus, once you click on to a particular VerticalNet site, you're automatically linked to our other related sites to give you a cross
pollination of ideas that just can't be beat. VerticalNet. Anything else...is old news.

Get your FREE Photo Mousepad, now!
Just log on to www.engineeronline.com

VerticalNet

Select No.20 WebAdvantage

Fig. 14. The final text extraction results.
6. Conclusions

◆ Text extraction method for complex color documents
  – Color quantization technique: ACR
    • Using suitable split and merging conditions
  – Identifying text regions: PLA technique
    • To improve the shapes of the characters
      – SOFM neural network
  – Merging procedure
◆ Advantages
  – Relatively robust to variations in font, color, or size