Joint Conference of 1996 International Computer Symposium
December 19~21, Kaohsiung, Taiwan, R.O.C.

SEGMENTATION AND UNDERSTANDING OF COLOR MAGAZINE IMAGES*

Shin-Ching Lin and Wen-Hsiang Tsai[†]

Department of Computer and Information Science
National Chiao Tung University
Hsinchu, Taiwan 300
Republic of China

Abstract

An integrated approach to segmentation and understanding of color magazines is proposed. First, a dynamic-cut color quantization method is proposed, which can be used to classify color image contents into more reasonable classes. Next, for block segmentation, a new algorithm is proposed to grow neighboring pixels with identical colors into blocks more efficiently. After block extraction, several features are used to classify extracted blocks into text blocks and graphic blocks. For construction of blocks from magazine images which usually have complicated background, improved methods for merging text blocks into text areas and graphic blocks into graphic areas are also proposed. These classified areas are then used to analyze the document layout. Methods for finding titles, abstracts, section titles, section texts, headers, footers, page numbers, and captions of graphics in magazines are proposed. Experimental results are shown to prove the feasibility of the proposed approach.

1. Introduction

In this paper, techniques for segmentation and understanding of color magazine images are proposed. A magazine image contains only raw data. It must be analyzed further before desired information can be gathered. The analysis works basically include image content segmentation, classification, and understanding.

The objective of the segmentation process is to partition an image into distinct regions of homogeneous regions. The methods can be broadly classified into three approaches, run length smoothing [1], projection profile cutting [2], and connected component analysis [3]. For color images, Lin and Tsai [4] proposed a multi-spectral constrained run length algorithm which can be used to segment blocks in parallel in all color planes. After segmentation of image contents into blocks, the next step is to recognize the block type. Chuang and Tsai [5] proposed some features to recognize the block type. The final step is document understanding. The objective is to extract the logical relationships among texts, graphics, and other components. The major works include structure recognition and layout understanding. Some related studies can be found in [6-10], most of which do not process magazine images.

When a color document is read in, the first stage of our approach is color quantization. The purpose of color quantization is to reduce the number of colors. A dynamic-cut color quantization algorithm is proposed in this study. Then, objects in the document image need be extracted. For this purpose, a new color block extraction algorithm is proposed. After blocks are extracted, line blocks and special components, including vertical lines, horizontal lines, tables, and frames then are extracted. The blocks which are not line blocks and special components are regarded as basic blocks. Several features are proposed to classify these basic blocks into graphic blocks and text blocks. Then, text blocks need be grouped into text lines and then into text areas. For this purpose, an algorithm, which avoids merging text blocks with background or graphic blocks erroneously, is proposed. On the other hand, overlapping graphic blocks are also merged to form graphic areas. After document segmentation, the layout of the document is "understood". By utilizing knowledge of the magazine layout, methods for extracting specific text or graphic regions, like titles, abstracts, section titles, section texts, page numbers, headers, footers, and

*This work was supported by National Science Council under Grant NSC85-2213-E-009-014.

[†] To whom all correspondence should be addressed.

section texts, page numbers, headers, footers, and graphics, etc., are proposed. On the other hand, methods for linking the contents of an article on different pages and for separating different articles on different pages are also proposed. Because of page limit, only some of the proposed methods will be described in the following sections.

2. Color Quantization

A method of color quantization is usually designed to recursively partition the color space into quantization cubes as shown in Figure 1. Existing partitioning methods include median-cut [4], mean-cut [4], or center-cut [11] methods, in which the cutting point is the median, mean, or center of the cube, respectively. Each of the cubes is then assigned a representative color. And all the colors within a cube are regarded to form a class and reproduced by a representative color value.

In some documents, texts and background can be of any color. Colors of texts may be classified to those of background. The proposed dynamic-cut quantization method can be employed to classify color pixels more effectively. By using the center-cut method, the cutting points usually lie near a peak of the distribution as shown in Figure 2(a). The pixels of similar colors will be separated into different classes. The proposed dynamic-cut quantization method tends to cut near the valley points to avoid separating the pixels of similar colors, as shown in Figure 2(b). Furthermore, as shown in Figure 3(a), if region R1 and region R2 are classified into a class, the original colors of the points in region R2 will be far from the mean color of the class after quantization because there are very few pixels in region R2. By the proposed dynamic-cut method, R1 is separated from R2.

The method for finding the cutting point for proposed dynamic-cut quantization is described in detail as follows (see Figure 3(b)). Assume the length of the entire range of color space is H.

Step 1: Apply the center-cut quantization method to find the center-cut point P_o in the entire color range.

Step 2: Form the search region L with its middle point being P_o and its width being H/3.

Step 3: Trace the possible cutting point in region L from left to right, and let the currently traced point in L be denoted as P. Also, compute the values of L1, L2, R1, and R2, where R1 and R2 are the color subranges to the left and to the right of point P, respectively, and N1 and N2 are the numbers of pixels in R1 and R2, respectively.

Step 4: If the number of pixels on the traced point P (i.e., if the number of pixels with colors identical to that specified by the traced point P) is smaller than 100, the value of $L1 \times N1 - L2 \times N2$ is recorded. The reason why the threshold 100 is set is to ensure that the optimal cutting point found finally is near the foot of a "mountain" or a valley as shown in Figure 2 (b).

Step 5: Go to Step 1 to trace the next point until all points in region L are traced.

Step 6: Pick out the point whose corresponding value of $|L1 \times N1 - L2 \times N2|$ is the smallest as the desired cutting point. The desired cutting point makes R1 with large number N1 of pixels has small region length L1.

3. Segmentation of Basic Blocks and Special components

In this study, we propose a new algorithm to extract color blocks more efficiently, even though the color information is used. The algorithm is described as follows, in which the neighborhood of a pixel is defined to be the 5 × 5 region with the pixel as the center. Also, a block list is maintained for recording temporarily extracted blocks.

Step 1: Input an image which has been quantized. Initially, there is no block in the block list.

Step 2: Get a pixel P from the input image from left to right, and then from top to bottom.

Step 3: Check in the neighborhood of pixel P, whether there exists a block of the same color as that of pixel P. During searching the neighboring block, if a checked block are far enough from pixel P in the vertical direction in the previously checked part of the input image, this block is merged into its neighboring block with the same color as it. If there is a neighboring block of the same color around pixel P, pixel P is merged into the found block. Otherwise, pixel P is regarded as the starting point of a new block.

Step 4: Go to Step 2 until all pixels in the input image are traced.

Step 5: Merge the blocks which are close and have the same colors.

Step 6: Stop with the blocks in the final block list as the desired extracted blocks.

4. Basic Block Recognition

Some statistical features used in this study are as follows, in which the first two features come from [5] and the other two are proposed in this study.

(1) The saturation degree of a block [5]:

$$C = \frac{\textit{number of pixels extracted as pixels of block}}{\textit{block length} \times \textit{block width}}$$

(2) The compactivity degree of a block [5]:

$$C = \frac{\textit{number of isolated pixels extracted as pixels of block}}{\textit{number of pixels extracted as pixels of block}}$$

(3) The number of horizontal runs and vertical runs of a block:

The pixels of a graphic block are usually distributed uniformly. Therefore, if there is a very large number of runs in a block, this block is usually a graphic block. According to our experimental results, if the number of satisfied runs is larger than 5 × (block width + block height), this block can be decided to be a graphic block.

(4) The number of pixels with the same colors as those of background in a block:

There is usually background noise. A block is regarded as a noise block if the number of pixels with the same colors as those of background in the block is larger than (1/3) × (number of block pixels).

Basic blocks are classified into text blocks, graphic block, or false blocks in this study. False blocks are the blocks whose types cannot be determined. After extracting a primitive block, we classify it in the following way. If the type of a block is determined in a step, we stop going to the next step.

Phase 1 : During the process of segmentation :

Step 1: Recognize the block by the line-width feature. If the feature value is too small, the block is recognized to be a graphic block and go to Step 8.

Step 2: Recognize the block by its size. If the block is too small, it is regarded as a false block and go to Step 8.

Step 3: Recognize the block by the saturation degree of the block. If the number is too large, it is regarded as a graphic block and go to Step 8.

Step 4: Recognize the block by the compactivity degree of the block. If the number is too large, it is regarded as a false block and go to Step 8.

Step 5: Recognize the block by the number of the horizontal runs and the vertical runs of the block. If the number is larger than 5 × (block width + block height), it is regarded as a graphic block and go to Step 8.

Step 6: Recognize the block by the number of pixels in the block with the same colors as those of the background. If the number is larger than (1/3) × (number of block pixels), it is regarded as a false block and go to Step 8.

Step 7: The block is regarded as a temporary text block.

Step 8: The block is recognized.

Phase 2 : During the process of area construction:

For a block which is regarded as a temporary text block in Phase 1, perform the following steps to make sure that this block is really a text block.

Step 1: Recognize the block by trying to cut the block into characters. If the block cannot be cut into distinct characters, it is changed to be a false block and go to Step 4. The block is cut with a projection method.

Step 2: Recognize the block by the colors of the block and the background. If the color of the block is the same as the background color of the block, it is regarded as a false block and go to Step 4.

Step 3: The block is regarded as a text block.

Step 4: The block is recognized.

5. Text and Graphic Area Construction

5.1 Merging of Neighboring Textline Blocks

The next step is to merge neighboring textline blocks to form larger textline blocks. Because the size of a punctuation mark is small, textline blocks are always broken by the positions of punctuation marks. For convenience of future works, these broken textline blocks must be merged at first. To accomplish this, given two textline blocks, check whether they satisfy the following conditions; if so, then they are merged.

1. The directions of the two textline blocks are the same.
2. The difference between the character widths of the two textline blocks is smaller than a threshold value. The threshold value is set to be a third of the maximum character width of the two textline blocks.
3. The distance between the two textline blocks is smaller than a threshold value. The threshold value is set to be the maximum character width of the two textline blocks.

5.2 Merging Textline Blocks into Text Areas

After the construction of textline blocks, they need be merged into text areas. The proposed method of constructing text areas includes two phases and is described as follows.

Step 1 : Input the textline blocks. Initially, there is no text area.

Step 2 : Get a textline block or a text area A from right to left for a vertical document and from top to down for a horizontal document.

Step 3 : If there is a textline block B on the left side of A, the following conditions are checked. If they are satisfied, then B is merged into A.

1. The distance D between A and B is smaller than a threshold value. The threshold value is set to be 1.2 × the character width of A.
2. The difference between the character width of A and that of B is smaller than a be a third of the maximum character width of A and B.

Step 4 : If there is a textline block C included in or overlapping A, we merge C into A. C must be the textline block separated from B by punctuation marks.

Step 5 : If there are text pixels which are not extracted as text blocks and are neighboring to A, merge these pixels into A. The method is to search the pixels with the same colors as those of A in a region R. If there is a sufficient large number of found pixels, combine these pixels into A. Otherwise, these found pixels are regarded as noise and are ignored. They cannot be combined to A.

Step 6 : Go to Step 1 until all textline blocks are checked.

5.3 Merge of Graphic Areas and Captions of Graphics

Overlapping graphic blocks are merged into graphic areas. After merging graphic areas, the next step is to search the captions of graphic areas. A general way is to search the surrounding text areas of graphic areas. But there might be texts which are not captions and are close to graphic areas. For this reason, if the number of textlines of a text area which is close to a graphic area is larger than three, this text area is not merged into the graphic area. Here, we assume that the number of textlines in a caption is smaller than three.

6. Article Arrangement Understanding

6.1 Article Arrangement Understanding for Magazines

In producing a magazine, many rules are used in the composition of the articles. Although lots of them are professional knowledge, we can obtain useful information from observation of magazine contents as well as analysis of human beings' reading habits. Some of the information found in this study is listed as follows.

1. In nowadays magazines, the main principle for arranging articles is usually to make it rectangular in shape.
2. Vertical Chinese magazines are arranged from top to bottom and then from right to left.
3. Horizontal Chinese magazines are arranged from left to right and then from top to bottom.
4. An article is always composed of an article title at the beginning, followed by its abstract and its text.
5. The author name of an article is usually neighboring to the article title.
6. The page number of a page is always on the center, left, or right of the bottom of the page.
7. The headers are at the top, and the footers are at the bottom, of a page.
8. The distances between section titles and section texts are usually larger than the distances between the text lines of section texts.
9. Normally, there exists a caption around a graphic.
10. The characters of article titles are generally larger in size than those of the text lines in section texts.
11. The characters of section titles are generally larger in size than those of the text lines in section texts.
12. The characters of the text lines in a section text usually have a uniform size.

There are multiple abstract, multiple sections, and multiple graphics regions in an article, and there are multiple section text regions in a section. We must record the abstract number, the section title number, the section text number, and the graphic number to achieve the aim of linking the contents of documents. When a type of region is extracted, the corresponding number of the region type is added one. Also, it is desired that the regions of a type can be extracted according to the human reading order. For this reason, the order of the region number is taken to be the human reading order. The process of understanding an article is described as follows.

Step 1 : Input the constructed text areas and graphic areas of a magazine image.

Step 2 : Find the article title region. If the size of the characters of a text area is larger enough, the abstract number, the section number, the section text number, and the graphic number are set to zero because a new article is just found and its content will be extracted next.

Step 3 : Find the author name region. If a text area satisfies the following criteria, it is regarded as an author name region.

1. The location of the text area is on the right-bottom side of a horizontal title region or on the left-bottom side of a vertical title region.
2. The text area has only one text line because an author name region has only one text line.
3. The number of characters is smaller than four because the largest number of characters of an author name region is four.

If the text area satisfies Criterion 1 and does not satisfy Criteria 2 and 3, it is regarded as an abstract region, not an author name region. The abstract number is incremented one and assigned to this abstract region. In case this situation occurs, it is checked further the existence of the author name region. If a text area satisfies the following criteria, it is an author name region.

1. The location of the text area is on the left-bottom side of the found abstract region with the vertical direction or on the right-bottom side of the found abstract region with the horizontal direction.
2. The text area has only one text line.
3. The number of characters is smaller than four.

Step 4 : Find the header regions. The text areas at the top of the document image are regarded as the header regions.

Step 5 : Find the page number region. If a text area satisfies the following criteria, it is regarded as a page number region.

1. The text area is on the left, center, or right side of the bottom.
2. There is only one text line.
3. The size of the characters is small.
4. The number of characters is smaller than five.

Step 6 : Find the footer regions. The text areas which are not page number regions and are at the bottom of the document image are regarded as footer regions.

Step 7 : Find the section title regions and the section text regions. The proposed procedure to find section title regions and section text regions is described as follows.

Step 7.1 : Sort text areas which are not understood before by the reading order. For vertical documents, areas are sorted from right to left and then from top to bottom. For horizontal documents, areas are rearranged from top to bottom and then from left to right.

Step 7.2 : Decide the size of characters and the distance of text lines for section regions. The text area which has the largest number of text lines is regarded as a section text region. The size of characters and the distance of text lines for section texts are those of this text area.

Step 7.3 : Decide the article orientation. The textline orientation of the section text region chosen in Step 7.2 is regarded as the article orientation.

Step 7.4 : Find the first section text region from the sorted text areas. If a text area satisfies the following criteria, it is regarded as a section text region.

1. The number of text lines of the text area is larger than three. Otherwise, it must be on the left border of the vertical-oriented document, or on the lower border of the horizontal-oriented document.
2. The size of the characters of the text area is identical to the one decided in Step 7.2.
3. The distance of the text lines of the text area is identical to the one decided in Step 7.2.

Step 7.5 : Find the section title region of the found section text region. If a text area satisfies all of the following criteria, it is regarded as the section title region of the found section text region.

1. The size of the characters of the text area is larger than the size of the found section text region.
2. The text area is just on the right of the found section text region for a vertical document, or on the top side of the found section text region for a horizontal document. Otherwise, the section title region might be on the left side of the upper column for a vertical document (see Figure 11(a)), or on the bottom side of the left column for a horizontal document (see Figure 11(b)).

After the section title region is found, the section title number is incremented and assigned to the found section title region, and the section text number are set to one and assigned to the found section title region.

Step 7.6 : If there is no section title region for the found section text region and there is no section text region extracted before extracting this found section text region, do the following steps.

Step 7.6.1: The found section text region is regarded as an abstract region. The abstract number is incremented one and assigned to this abstract region (see Figure 12).

Step 7.6.2: Removed this abstract region from the sorted text areas.

Step 7.6.3: Go to Step 7.3.

Step 7.7 : The section text number is incremented and assigned to the found section text region. Remove the section title region and the section text region found from the sorted text areas.

Step 7.8 : Go to Step 7.4 until no more text area can be found in the sorted text areas.

Step 8 : Text areas which are not understood so far are regarded as abstract regions if their numbers of characters are larger than three and they do not overlap the understood regions.

Step 9 : Text areas which are not understood so far are regarded as noise if their numbers of characters are smaller than or equal to three and they overlap the extracted text areas.

Step 10 : Sort graphic areas in the way as described in Step 7.1 and extract graphic regions by the order of the sorted text areas. The graphic number is incremented and assigned to extracted graphic regions.

Step 11 : Now, all text regions and graphic regions in the magazine image have been understood. There might exist the next page for understanding.

6.2 Understanding of Multi-Page Articles

When understanding a magazine image, the abstract number, the section number, the section text number, and the graphic number for all regions are recorded and the order of these numbers matches the human reading order. During understanding of the next page, all regions are also extracted by the reading order and the numbers will be incremented, following those of the last page, according to the types of regions found. In this way, the contents in neighboring pages are linked by these numbers. Furthermore, if a title is found in the current page during analyzing several pages all types of numbers are set to zero as a preparatory step to start processing the new article.

A graphic region may appear on two neighboring pages. The two parts of a graphic region need be regarded as a single graphic region and assigned an identical graphic number. The method for vertical documents is described as follows. The method of merging graphic regions in neighboring pages for horizontal documents is similar to the one for vertical documents.

For vertical documents, the graphic regions which touches the left edge of a page are recorded. For a graphic region of the next page, if it satisfies the following criteria, it is regarded as part of the graphic region on the current page.

1. The graphic region which touches the right edge of the current page.
2. The upper boundaries of the graphic areas in the two pages are quite close, or more specifically, $|Y1-Y2|<10$, where Y1 is the Y-coordinate value of the upper boundary of the graphic area in the current page, and Y2 is the Y-coordinate value of the upper boundary of the graphic area in the next page.
3. The upper boundaries of the graphic areas in the two pages are quite close, or more specifically, $|T1-T2|<10$, where T1 is the Y-coordinate value of the lower boundary of the graphic area in the current page, respectively, and T2 is the Y-coordinate value of the lower boundary of the graphic area in the next page, respectively.

7. Experimental Results

Several images obtained from magazines were tested by the proposed approach on a DX4-100 PC using the C++ language. The tested images were obtained by an Umax vista s-8 color scanner at 200 dpi. Some segmentation results are shown in Figure 4. The results for extracting blocks can be seen to be good. The contents in multiple pages are linked by the block number. The average execution time of each processing step is shown in Table 1 for an image with size 1250 x 1650.

Table 1 The execution time of each processing step.

Processing step	time
dynamic-cut color quantization	*2 sec*
Extraction of special components, text, and graphic areas	*110 sec*
Article arrangement understanding	*5 sec*

8. Conclusions

An image analysis system including segmentation and understanding for color magazines has been successfully implemented. Several major achievements in different processing stages are summarized as follows.

In the stage of color quantization, a dynamic-cut color quantization method has been proposed. In the stage of block extraction, a new color block extraction algorithm has been proposed. Text blocks and graphic blocks are classified using a recognition method which combines the use of statistical features, structural features, and several processes. In the stage of text area construction, methods for merging false blocks and merging basic text blocks into text area blocks have been proposed. In the article arrangement understanding phase, a systematic procedure to understand the extracted blocks in magazine images has been proposed. Some arrangement rules and reading habits of human begins for magazines were used to extract the relationships among segmented blocks and graphics, and those among multiple articles in multiple pages. The experimental results have revealed the feasibility of the above proposed algorithms.

References

[1] F. M. Wahl, K. Y. Wong, and R. G. Casey, "Block Segmentation and Text Extraction in Mixed Text/Image Documents," *Computer Graphics and Image Processing 20*, pp. 375-390, 1982.

[2] D. Wang and S. N. Srihari, "Classification of newspaper image blocks using texture analysis," *Computer Vision, Graph, and Image Process*, Vol. 47, 1989, pp. 327-352.

[3] L. A. Fletcher and R. Kasturi, "A robust algorithm for text string separation from mixed text/graphics images," *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 10, 1988, pp. 910-918.

[4] Y. S. Lin and W. H. Tsai, " Image Segmentation for color Document Analysis," in *Proc. Int. Conf. Computer Vision, Graphics and Image Processing*, Taoyuan, Taiwan, Republic of China, August 1994, pp. 135-142.

[5] Y. H. Chuang and W. H. Tsai, "Segmentation of Texts, Graphics, and Special Components for Color Document Image Analysis," in *Proc. Int. Conf. Computer Vision, Graphics and Image Processing*, Taoyuan, Taiwan, Republic of China, August 1995, pp. 471-478.

[7] J. Toyoda, Y. Noguchi and Y. Nishimura, "Study of Extracting Japanese Newspaper Article," *Proc. 6th ICPR*, Munich, pp. 1113-1115, 1982.

[8] F. Esposito, D. Malerba, and G. Semeraro, "An Experimental Page Recognition System for Office Document Automatic Classification: An Integrated Approach for Inductive Generalization," *Proc. 10th Int. Conf. Pattern Recognition*, Atlantic City, NJ, USA, June 1990, pp. 557-

562..

[9] K. K. Lau and C. H. Leung, "Layout Analysis and Segmentation of Chinese Newspaper Articles," *Computer Processing of Chinese and Oriental Languages,* Vol. 8, No. 1, pp. 97-114, 1994.

[10] L. F. Lee and W. H. Tsai, "Understanding of Arrangments and Extraction of Articles in Chinese Newspaper Images," in *Proc. Int. Conf. Computer Vision, Graphics and Image Processing,* Nantou, Taiwan, Republic of China, August 1995, pp. 479-487.

[11] G. Joy and Zhigang Xiang, "Center-cut for color-image quantization," Visual Computer, Vol. 10, 1993, pp. 62-68.

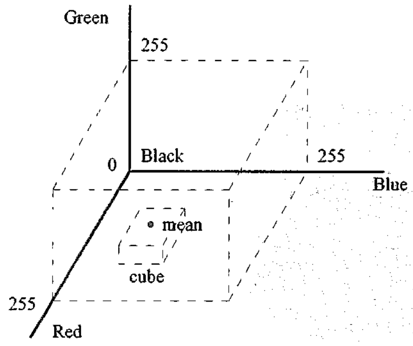


Figure 1: A quantized cube associated with a representative value.

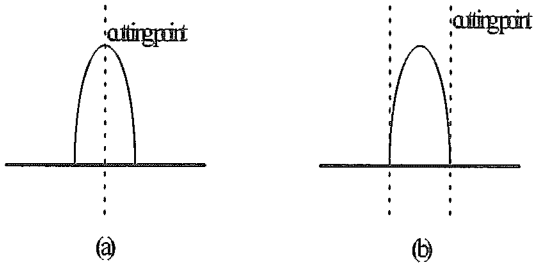
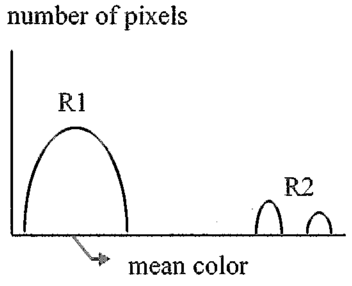
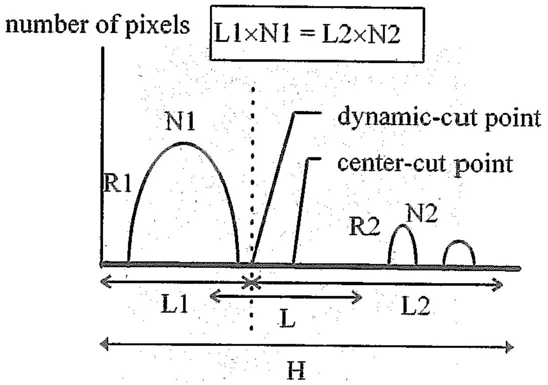


Figure 2: Comparasion of two methods for color quantization. (a) Classification by center-cut color quantization. (b) Classification by proposed dynamic-cut color quantization

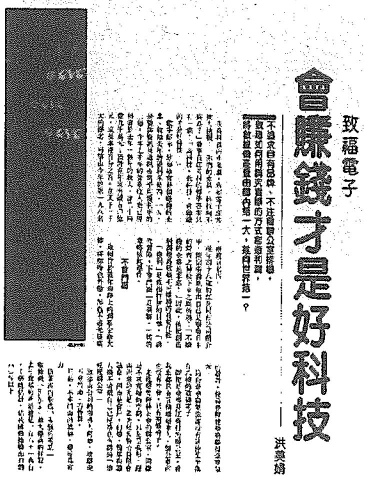


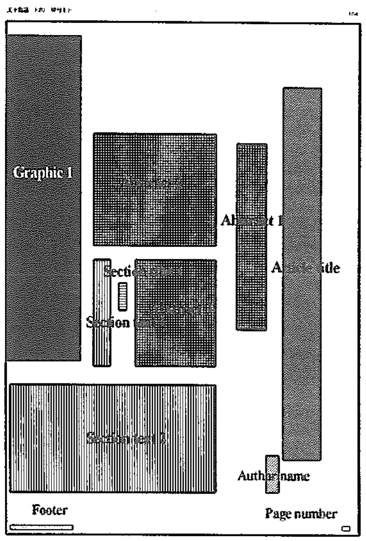
(a)



(b)

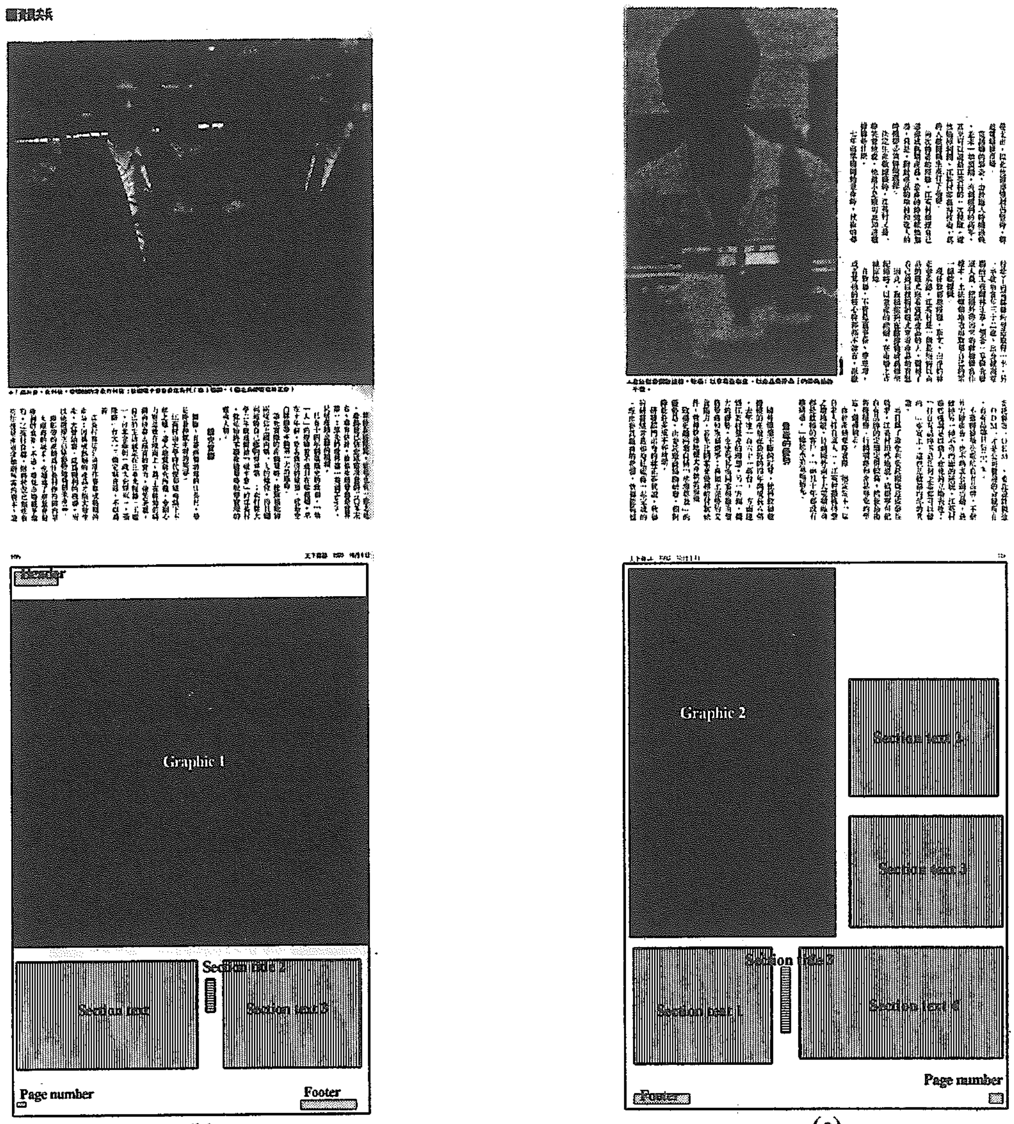
Figure 3 Illustration of two methods for color quantization. (a) Classification by center-cut color quantization. (b) Classification by dynamic-cut color quantization. N1 and N2 are the numbers of pixels of R1 and R2, respectively; L1 and L2 are the lengths of R1 and R2; respectively, and L is the region whose center point is the center-cut point and whose length is H/3.


致福電子
會賺錢才是好科技
洪美娟



(a)

Figure 4: Experimental results for a vertical article. (a) The first page. (b) The second page. (c) The third page.



(b)

(c)

Figure 4: Experimental results for a vertical article. (a) The first page. (b) The second page. (c) The third page(continued).