REGION AND KEYWORD EXTRACTION BASED ON COLOR MARKING FOR DOCUMENT ENTRY

Mineo Shoman† , Takashi Nishimura‡ , Toshiya Kawauchi†

† NTT Human Interface Laboratories
1-2356, Take, Yokosuka-shi
Kanagawa 238-03, Japan

‡ NTT America, Inc.
4962 El Camino Real, Suite #230
Los Altos, CA 94022, U.S.A.

ABSTRACT

A method for extracting an article and its keywords from a notated document using color image processing is presented. In this method, the user selects a region of an article and its keywords and marks them in color on the printed document. The system extracts the article and recognizes the keywords. This paper describes, (1) colored area extraction and color effect removal by adaptive thresholding using pseudo-HVC coordinates based on the "transparent coloration model" which based on the behavior of colored pixels in color space, (2) polygonal shape extraction which extracts the intended article from a handdrawn outline by a new and simple algorithm "recursive vertex search", (3) and some experimental results which show high accuracy for newspaper article extraction using colored pixels in color space. The color marks of article borders and keywords are identified and then removed by adaptive thresholding using the pseudo-HVC coordinates. Because the color marks are handdrawn they will not be completely accurate. We have developed a polygonal shape extraction algorithm that examines the marked outline and then determines which part of the text is to be extracted. This algorithm is not yet complete because it needs information on typical document architecture.

When complete the system will extract the desired article and stored as an image in an electronic filing system. Keywords will be identified and used as the reference keys by which the article can be retrieved. These actions are not possible at this time and maybe the subject of further research. Experimental results from the processing of newspaper articles are presented, they confirm the accuracy of color mark identification and subsequent removal of color effects.

INTRODUCTION

In modern information age the volume of printed material is dramatically increasing. To permit rapid extraction and filing of useful articles an automatic process, that is easy to initiate is needed. Previously, such a system was proposed[1]; however, it employed a monochrome scanner which restricted its performance. We have developed a new process that uses color instead of monochrome. Performance and range of functions have been significantly improved. The extraction and filing of articles requires two basic steps: the identification of article borders, and the determination of keywords against which the article is filed. It is possible to manually input each article but this is considered too time consuming, and thus, too expensive. The most efficient way to process a printed text is to delineate the article with a colored line, keywords are underlined with a different color. These actions are very fast and performed off-line, they do not require computer usage. The marked text is scanned in color and the entire page is entered for processing. Final system configuration is shown in Fig.1.

The colored pixels are examined with a coloration model which is based on the activity of the behavior of

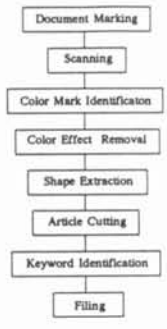


Fig.1 System Configuration

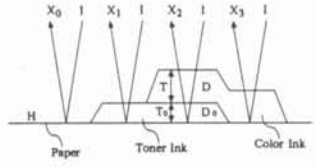


Fig. 2 Coloration Model

COLOR EXTRACTION AND COLOR EFFECT REMOVAL BASED ON COLORATION MODEL

We propose the coloration model to assess the reflectance from paper which is toned and/or colored for annotation. The model on which this method is based is shown in Fig.2. In this model, coloration by toner ink is assumed to be similar to that of color ink. The coloration is created by a film overlaying the paper. The film is assumed to have a certain transmittance and thickness, and has no inner or boundary reflectance. The transmittance is in proportion

to a power of its thickness, and has RGB components. The differences between transmittance components causes the perceived colorfulness. Thickness changes exist at the edges of color ink or toner ink. The color scanner observes the reflection intensity as a color level. Let the transmittance of color ink film per unit thickness be $D(D_R, D_G, D_B)$ and its thickness be T, the transmittance of toner ink film for the unit thickness be $D_0(D_{0R}, D_{0G}, D_{0B})$ and its thickness be T_0, the intensity of the light-source be $I(I_R, I_G, I_B)$, and the reflectance of the surface on the paper be $H(H_R, H_G, H_B)$. Let the intensity of the reflection on the paper be $X_0(X_{0R}, X_{0G}, X_{0B})$, that through toner ink be $X_1(X_{1R}, X_{1G}, X_{1B})$, that through color ink and toner ink $X_2(X_{2R}, X_{2G}, X_{2B})$, and that through color ink only be $X_3(X_{3R}, X_{3G}, X_{3B})$.

$X_{0P} = H_P I_P$ Where, $P = R, G, B$ and so on.

$X_{2P} = D_P^T D_{0P}^{T_0} H_P D_{0P}^{T_0} D_P^T I_P$

So;

$X_{2P} = X_{0P} D_{0P}^{2T_0} D_P^{2T}$

On a log scale,

$x_2 = x_0 + t_0 d_0 + td$

Where, $x_{0P} = log X_{0P}$, $x_{2P} = log X_{2P}$, $d_P = log D_P$, $d_{0P} = log D_{0P}$, $t = 2T$, $t_0 = 2T_0$.

In log color space, toned and/or colored pixels exist on the plane including the paper color level and are identified by the toner vector d_0, and color vector d shown in Fig.3. Similarly;

$x_1 = x_0 + t_0 d_0$

Where, $x_{1P} = log X_{1P}$.

The color level of a pixel colored by the toner ink exists on the half line from the color level of paper which has the toner vector direction. So,

$x_2 = x_1 + td$

The color level of a pixel colored by the color ink exists on the half line which has the color vector direction from the color level prior to color application. Assume the color levels are projected on the color ink half line in the direction of toner vector, the resulting color levels are those without toner. The distance in log color space from the paper on the color ink half line, is in proportion to the thickness of color ink film. So, a pixel is classified colored or not by the threshold on the color ink half line between x_0 and x_3, where, $x_{3P} = log X_{3P}$.

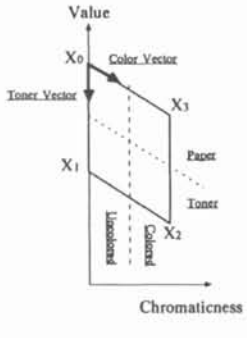


Fig. 3 Coloration Plane

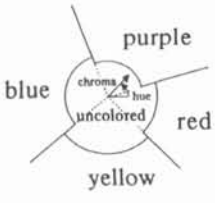


Fig. 4 Color Classification

This is the principle of colored pixel extraction, or colored area extraction. Conversely, suppose the color levels are projected on the toner ink half line in the direction of color vector, the result color levels are those of before application of color ink. The distance in log color space from the paper on the toner ink half line, is in proportion to the thickness of toner ink film. So, a pixel is classified toned or not by the threshold on the toner ink half line between x_0 and x_1. This is the principle of toned pixel extraction, or the removal of color effect.

In an actual process, the colored area extraction and color effect removal are performed in the cylindrical coordinates or pseudo-HVC space. Let the origin be x_0, that is the color level of the paper, and the z-axis be the direction of toner vector d_0. Then the z-axis corresponds to the Value but in the negative direction, the r-axis to Chromaticness and the theta-axis to Hue. Whether a pixel is colored or not is classified by chromaticness, and its color is classified by hue. An example of these thresholds are shown in Fig.4. Whether the pixel is toned or not is classified by the Value threshold which is modified by its chromaticness, shown in Fig.3.

Fig.5 is the distribution of color levels which are mapped on the plane corresponding to Fig.3. Fig.5(a) is the distribution of red ballpointpen on newspaper, and Fig.5(b) is that of red markerpen on newspaper. These distributions are distorted parallelograms, because the film thicknesses seem to be thinner on toner ink than on paper.

(a) Ballpointpen (red) on Newspaper

(b) Markerpen (red) on Newspaper

Fig.5 Color Level Distribution on Coloration Plane

SHAPE EXTRACTION METHOD

This chapter describes the method for extracting the shape of the intended region from the color plane. As the first approximation, rectangles are created each of which encloses one colored area consisting of connected colored pixels. This system uses the method based on the projection profile value[1]. The method is relatively strong against noise and needs less processing time. From the color plane of the keywords color, the extracted rectangles are the shape of the keyword regions, because keywords are rectangular. The method of extracting the article shape from its colored outline is called the "recursive vertex search"[2], and is shown in Fig.6. It approximates the shape with a polygon

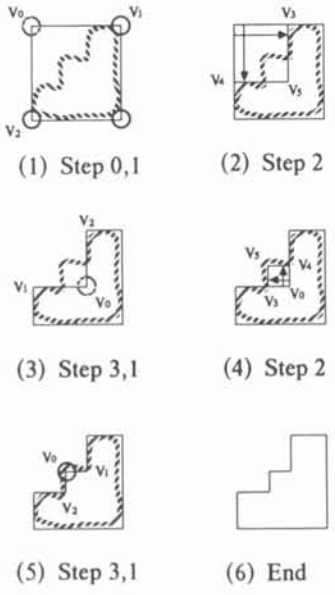


Fig. 6 Recursive Vertex Search

Fig.7 Sample Data

Fig.8 Restored Data

which has vertical and horizontal edges. Most articles have vertical and horizontal boundaries. The process is as follows; For each of the four vertices of the rectangle;

(Step0) The concerned vertex is set to $V_0(X_0, Y_0)$, the vertex which is adjacent to V_0 in the x-axis direction (horizontal) is set to $V_1(x_1, y_0)$, and in the y-axis direction (vertical) is set to $V_2(x_0, y_1)$.

(Step1) If a colored outline exists in the neighborhood of V_0, process is stopped, if not go to Step2.

(Step2) The colored outline which is first found along the edge from V_0 to V_1 is located, its x-coordinate is set to x_2, and the point (x_2, y_0) is called V_3. Similarly V_0 to V_2 is searched, and y-coordinate of V_4 measured, the coordinates of V_4 are (x_0, y_2). The point V_5 is created at (x_2, y_2).

(Step3) The edges $V_5 - V_3$ and $V_5 - V_4$ are created, and $V_0 - V_3$ and $V_0 - V_4$ are deleted. V_0 is replaced with V_5, and V_1 with V_4, V_2 with V_3. Then Step1 is repeated.

This method cannot prevent the inclusion of regions which intrude into the article.

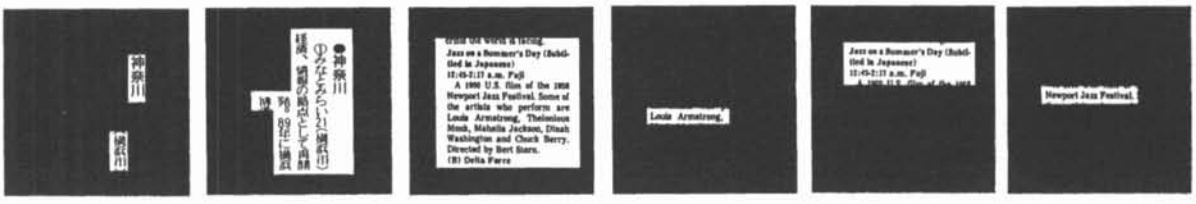
EXPERIMENTAL RESULTS

To confirm the accuracy of the system 40 articles from 8 newspapers were marked with a variety of colors and ink types. Prior to this confirmation the thresholds, except for Value scale, were determined from color and toner calibration vectors. These were prepared using combinations of 8 different newspapers, printed in Japan, marked with 2 ballpointpens (red and blue) and 4 markerpens (red, purple, blue and yellow). Each type of pen was made by one company. Ballpointpens and markerpens were not used in combination, because the reds overlap each other in hue.

Fig.7(a) shows a typical annotated article marked with ballpointpens, the article color is blue, and keywords are marked in red. In Fig.7(b) another article is shown marked with markerpens, the article colors are yellow and red, and keywords are colored blue and purple. Because this paper is monochrome, some explanation is necessary. Fig.7(a) and (b) shows the original articles as scanned by a monochrome scanner and binarized by the best thresholds for uncolored areas. Both of them have color effects. In Fig.7(a) ballpointpen lines are evident. In Fig.7(b) annotated keywords are a little blurred. These are "Newport Jazz Festival" and "Louis Armstrong" in the text. It seems that the color effects are insignificant in these samples. However, because thresholding is very critical of noise, the color effects significantly retard reading or character recognition. The original articles were scanned with a color scanner as multileveled color images. Fig.s 8 (a) and (b) are the binarized images of the original articles from which the color effects have been removed. The binarizing thresholds were selected to get images equivalent to respective images shown in Fig.s 7 (a) and (b). Fig.s 9 (a)-(f) show detected colored pixels. Where the colored area lies over toner the system often fails to detect the coloration and hence extracted colored lines may not be solid. It is thought that the toner ink has the effect of making the color ink film thinner, this renders the model invalid. However, this effect is not serious and accurate shape extraction is still possible. Fig.s 10 (a)-(f) show the shape of extracted regions for articles and keywords.

(a) red(Ballpointpen) (b) blue(Ballpointpen) (c) yellow(Markerpen) (d) purple(Markerpen) (e) red(Markerpen) (f) blue(Markerpen)

Fig.9 Detected Color Pixel



(a) red(Ballpointpen) (b) blue(Ballpointpen) (c) yellow(Markerpen) (d) purple(Markerpen) (e) red(Markerpen) (f) blue(Markerpen)

Fig.10 Extracted Image

Recog.	スペクトル
2nd Cand.	エベタト舟
3rd Cand.	ユベ゛←A
4th Cand.	只べ゛h井

Recog.	周期信号
2nd Cand.	届朋儀培
3rd Cand.	屈捌儒母
4th Cand.	用朔眉署

Fig.11 Keyword Recognition

The results are faithful to the original colored areas. Because this system does not know of typical document architectures, the extracted areas do not follow the actual article layout.

For 40 colored areas in 8 newspapers the shape of the extracted colored areas perfectly matched the marked areas. These newspapers were different from the ones used for determining the thresholds.

Fig.11 shows a sample of recognized keywords. Character recognition[3] works well because color effects have been removed.

FURTHER RESEARCH TOPICS

To improve this system the following items must be considered:

- knowledge of document architecture must be utilized in the shape extraction stage to ensure articles are accurately extracted.
- improved automatic color and toner vector detection is needed to accommodate a wide range of papers and pens.
- easier area marking by using simpler cutout symbols, such as marking just the article's corners.

CONCLUSION

A method is presented which extracts articles and their keywords from a notated document using color image processing. It is confirmed that colored area extraction, the color effect removal based on "transparent coloration model", and the polygonal shape extraction "recursive vertex search" work with high accuracy. Character recognition works well when color effects are removed. This system is applicable for automated entry systems, and should prove to be extremely useful.

ACKNOWLEDGMENTS

The authors would like to thank Messrs. I. Kawashima, T. Kawatani, K. Ishikawa, S. Tada, M. Hase, T. Akiyama and T. Fujimoto for their assistance and encouragement.

REFERENCES

[1] M.Hase, et al.: A Method for Extracting Marked Regions from Document Images, Proc. 8th ICPR p780, 1986

[2] T.Nishimura, et al.: A Method for Extracting Regions and Keywords from Color Marked Document (in Japanese), The Journal of Institute of Image Electronics of Japan, Vol.17, No.5, Oct.1988 (this will be published in October 1988)

[3] S.Tada, et al.: A Small Parallel Processor and its Application for Character Recognition (in Japanese), Trans. IEICE Jpn., J71-D, No.8, p1546, 1988