A ROBUST STATISTIC METHOD FOR CLASSIFYING COLOR POLARITY OF VIDEO TEXT

Jiqiang Song, Min Cai and Michael R. Lyu
Dept. Computer Science & Engineering, The Chinese University of Hong Kong, Hong Kong, China
{jqsong, mcai, lyu}@cse.cuhk.edu.hk

ABSTRACT

Video text extraction and recognition are prerequisite tasks for video indexing and retrieval. Color polarity classification of video text is very important to these tasks. Most existing text extraction methods assume that the text color is always light (or dark). Obviously, this assumption restricts the application of these methods to some specific domains. Only a few methods can detect the color polarity on condition that the background is clear. However, many real video texts have various appearances and complex backgrounds that existing methods cannot handle. This paper proposes a statistic color polarity classification method that is robust to various background complexities, font styles, stroke widths, and languages. We discover the intrinsic relationships between text edges and background edges, and then develop an efficient measurement to detect the color polarity. The experimental results show that the proposed method achieves a much higher accuracy, 98.5%, than existing methods.

1. INTRODUCTION

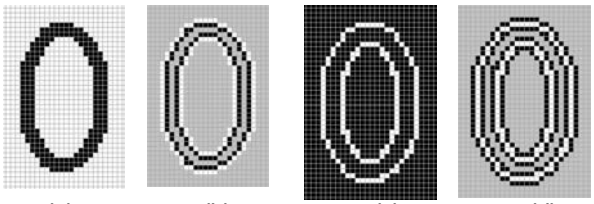
With the vast amount of available video sources, emerging techniques for video indexing and retrieval are of timely importance and interest. Video text, especially the super-imposed artificial text, is no doubt the most important clue for these purposes. Since texts may be assigned different colors, knowing the color polarity of video text (i.e. a text is light or dark) is very important for the correct video text extraction and recognition. Antani *et al* [1] proposed a text extraction method that can detect the color polarity. It firstly binarizes the text image into a positive image and a negative image. Then, for each image, a connected-component analysis is performed to filter out very small or large components and components that do not have aspect-ratio characteristics of text. A score is assigned to each polar image based on its text-like characteristics. Finally, the image with the higher score is selected and the other is discarded. This method works well for a text on a clear and contrastive background (Fig. 1-a); however, it cannot handle complex backgrounds (Fig. 1-b, 1-c) and various text appearances (Fig.1-d) since it depends heavily on the thresholding results. The failure of Fig. 1-d is because the black contours of texts also form many text-like connect components. In fact, the evaluation of text-like characteristics has to consider different font styles, stroke widths, and even languages. Unfortunately, these constraints may conflict with each other. Since a robust text color polarity classification method is not available yet, most existing text extraction methods [2,3,4] hold an assumption that the text color is always light (or dark). Obviously, this assumption restricts the application of these methods to some specific domains, e.g. captions.

This paper proposes a statistical text color polarity classification method that is robust to various background complexities and text appearances. It explores the intrinsic relationships between text edges and background edges, and designs an efficient measurement to detect the color polarity. The experiments show the proposed method is very accurate.

2. ANALYSIS

In general, the text color (or the background color) in a text image can be classified into *White* (*W*) and *Black* (*B*) using a proper threshold. Accordingly, there are four possible combinations (*XonY*) of the text color (*X*) and the background color (*Y*): *WonB*, *WonW*, *BonW*, and *BonB*. *WonB* and *BonW* are normal cases. For the other two cases, text strokes are usually surrounded by contrastive

Figure 1. Various text appearances and backgrounds



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

Figure 2. Original images (a), (c) and edges (b), (d)

contours to be distinguished.

A robust color polarity classification method must rely on a feature independent of background complexities, font styles, stroke widths and languages. Edge is eligible since a text must have a certain contrast against its background. We found that, in a binarized text image, the number of white edges (N_w) and the number of black edges (N_b) have intrinsic relationships with the text polarity. For example, Figure 2-a and 2-c show a black character 'O' on a white background (the *BonW* case) and on a black background (the *BonB* case), respectively. Their edges are shown in Fig. 2-b and 2-d, respectively, where white edges are in white, black edges in black and non-edges in gray. In Fig. 2-b, N_b is similar to N_w, while in Fig. 2-d, N_b is nearly twice as N_w. For the *WonB* case and the *WonW* case, we can find analogical relationships. More characters in a text string, more reliable this relationship is. Therefore, we can utilize this statistic property to detect text color polarity.

3. ALGORITHM

The proposed algorithm takes a grayscale text image as input. The output color polarity is either *Black* or *White*. Its seven steps will be described in the processing order.

3.1. Thresholding

For binarization, we employ the Otsu method [5] to find the best threshold according to the grayscale histogram of the text image. In the binary image, the value of a black pixel is 0 and that of a white pixel is 1.

3.2. Edge detection

All pixels in the binary image are classified into black edges, white edges and non-edges. We use two 3×3 convolution kernels [6] to detect edges (Fig. 3).

$$K_w(x,y) = \begin{bmatrix} 0 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 0 \end{bmatrix} \qquad K_b(x,y) = \begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

Figure 3. Edge detection kernels

$K_w(x,y)$ is for white edges, while $K_b(x,y)$ is for black edges. A pixel (x,y) is classified as follows:

$$P(x,y) = \begin{cases} White\_Edge, & K_w(x,y) > 0 \\ Black\_Edge, & K_b(x,y) > 0 \\ Non\_Edge, & K_w(x,y) \le 0 \text{ and } K_b(x,y) \le 0 \end{cases}$$

3.3. First ratio calculation

After the edge detection, we obtain the number of white edges and that of black edges, i.e. N_w and N_b. We then define the ratio of them as the first ratio,

$$R_1 = \frac{N_w}{N_b}.$$

R_1 can distinguish the *WonW* case from the *BonB* case, since the number of background edges is surely more than the number of text edges but less than twice of it. For the former, R_1 is larger than 1 and smaller than 2, while for the latter, R_1 is smaller than 1 and larger than 0.5. However, R_1s for both *WonB* and *BonW* cases are near 1. Therefore, R_1 itself is not sufficient for the classification.

3.4. Outmost edge removal

The edge distributions inside characters are too complex to analyze due to unknown stroke structures. However, for each character or a group of connected characters, it is safe to say that most of the outmost edges belong to the background color. Therefore, if the outmost edges are removed correctly, the change of N_w and N_b will be a good indication to the color polarity.

Since there is no guarantee that the edge of a character is closed, the contour following method cannot be used to remove the outmost edges. Instead, we remove them by the inward scanning (Fig. 4).

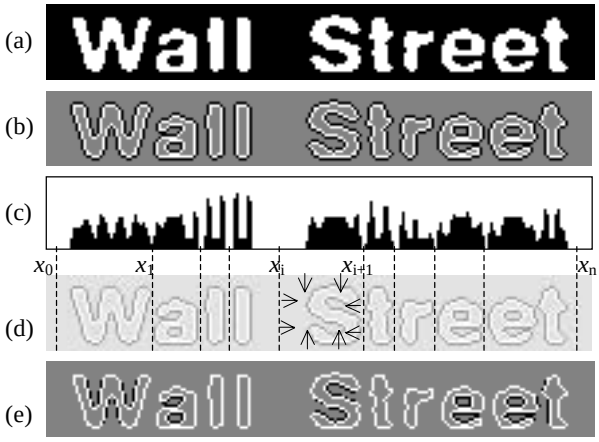


Figure 4. Inward scanning

Fig. 4-b shows the edge map of a binarized text image (Fig. 4-a), whose size is *width*×*height*. Non-edge pixels are in gray color. Firstly, the edge image is horizontally divided by $\{x_0, x_1, \ldots, x_n\}$, where $x_i \in [0, width]$ is located from the vertical projection of the edge map (Fig. 4-c). The midpoint of each absolute valley becomes x_i.

Then, in each rectangle $Rect_i$ (i=0..n-1) constructed by $y=0$, $y=height$, $x=x_i$, and $x=x_{i+1}$, we perform the inward scanning from four sides, as shown in Fig. 4-d. In each scan line, the first encountered edge pixel is removed.

After all rectangles have been processed, most outmost edges are removed (Fig. 4-e).

3.5. Second ratio calculation

The current number of white edges and that of black edges are denoted by N_w' and N_b', respectively. We define the ratio of them as the second ratio,

$$R_2 = \frac{N_w'}{N_b'}.$$

Now, R_2 can distinguish the *WonB* case from the *BonW* case. For the former, most of the removed outmost edges are black; therefore, R_2 increases from R_1 to be significantly larger than 1. Contrarily, R_2 decreases from R_1 to be significantly smaller than 1 for the latter. For the other two cases, the changes from R_1 to R_2 also provide more evidences for the classification.

3.6. Ratio variation normalization

If the background color is pure, the four cases can be classified by the two-level classification of R_1 and R_2, as shown in the first two rows in Table 1.

Table 1. Ratio variations of four cases

Case / Item	*BonW*	*WonW*	*BonB*	*WonB*
R_1	$R_1 \cong 1$	$1< R_1 \le 2$	$0.5 \le R_1 < 1$	$R_1 \cong 1$
R_2	$R_2 << 1$	$R_2 \cong 1$	$R_2 \cong 1$	$R_2 >> 1$
ΔR	Large –	Small –	Small +	Large +

But, in real video, the backgrounds of many texts are complex, i.e. containing both white color and black color. Considering the complex background, it is not easy to find an absolute threshold between "larger than 1"and "near 1". Another difficulty is that R_2 is unbounded. However, the change from R_1 to R_2 is a relative value; therefore, it is more reliable. We thus define the normalized ratio variation as follows:

$$\Delta R = \frac{R_2 - R_1}{\max(R_1, R_2)}, \qquad -1 \le \Delta R \le 1.$$

The theoretical estimation of ΔR for the four cases is listed in the last row of Table 1. Due to the background complexity, we cannot expect a clear separation between adjacent cases in the 1D ΔR dimension. To clear the ambiguities at the borders of two adjacent cases in the ΔR dimension, i.e., "Large –" to "Small –", "Small –" to "Small +", and "Small +" to "Large +", R_1 is taken into account since it is bounded. We expect that the four cases have separate distributions in the R_1-to-ΔR space ($0<R_1 \le 2$, $-1 \le \Delta R \le 1$). We select 200 training text images to check their distributions. The result is shown in Figure 5.

3.7. Color polarity classification

Figure 5 confirms that the four cases are separated in the 2D R_1-to-ΔR space. There is a clear gap between the *BonW* case and the *WonW* case. Since they are slightly overlapped in horizontal direction, we use a pair of thresholds (T_{1l} =-0.25, T_{lh} =-0.15) to classify them. In the

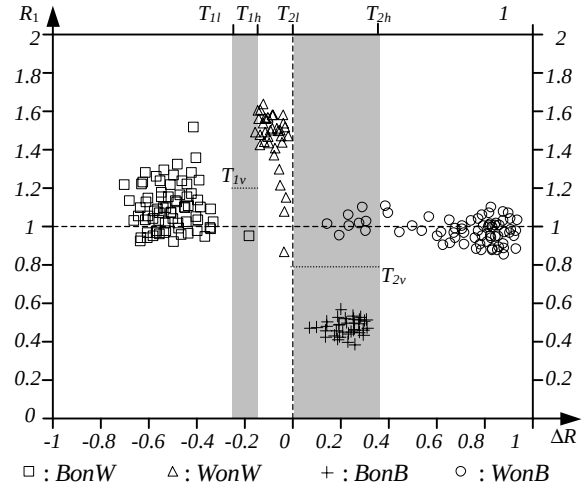


Figure 5. Distribution of four cases in the R_1-to-ΔR

critical part between T_{1l} and T_{lh}, a vertical threshold (T_{1v} =1.2) can distinguish them clearly.

The line ΔR=0 separates the *WonW* case from the right two cases completely. But, the *BonB* case and the *WonB* case have a relatively wide overlapping part. We use another pair of thresholds (T_{2l} =0, T_{2h} =0.35) to classify these three cases. T_{2l} can distinguish the *WonW* case from the other two cases. In the critical part between T_{2l} and T_{2h}, a vertical threshold (T_{2v} =0.8) can separate the *BonB* case from the *WonB* case clearly.

One may question that the *WonB* case tends to reach the *WonW* case. This may happen when the background contains many noises. However, we do not treat it as a problem since the misclassification between the *WonW* case and the *WonB* case does not affect the output of color polarity. They both output the *White* polarity.

Based on the above analysis and the thresholds obtained from the training data, the color polarity classification can be implemented efficiently as Figure 6.

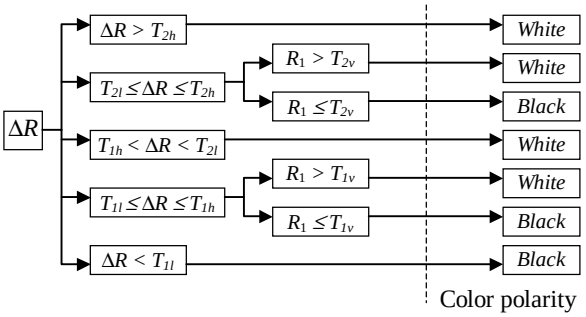


Figure 6. Color polarity classification

4. EXPERIMENTAL RESULTS

To test the accuracy of the proposed text color polarity classification method, we collect another 200 text images as the testing data. These images include different combinations of text color and background color, various

Figure 7. Part of the testing data

font sizes and styles, different stroke widths, and multilingual texts. Figure 7 shows some of these text images to demonstrate the coverage of the testing data.

The proposed method only generates 3 classification errors for 200 testing images, i.e., the accuracy is 98.5%. The classification speed is also fast. The average time for classifying a text image is 0.012 seconds (PIII500 CPU).

We also test the classification method proposed in [1] for comparison. Since they didn't indicate the binarization algorithm, the Otsu thresholding method [5] is employed. This method generates 36 classification errors and 7 reject errors for 200 testing images, i.e. the accuracy is 78.5%. The reject error happens when the score for the positive image and that for the negative image reach a tie. Most classification errors happen when the background is complex or the combination is the *WonW* case. The average speed is 0.027 seconds per image.

Therefore, the experimental results confirm that the proposed method is very robust to various background complexities and text appearances.

5. CONCLUSIONS

Text color polarity classification is the prior, however, crucial task to the video text extraction and recognition. This paper proposes a novel color polarity classification method based on the robust statistic relationships between text edges and background edges. The experimental results demonstrate that the proposed method is both accurate and fast. Furthermore, the proposed method can be seamlessly combined with those text extraction methods that hold a color polarity assumption, so as to extend their capabilities to handle various background complexities and text appearances.

6. ACKNOWLEDGEMENT

The work described in this paper was fully supported by two grants from the Hong Kong Special Administrative Region: the Hong Kong Research Grants Council under Project No. CUHK4360/02E, and Innovation and Technology Fund, under Project No. ITS/29/00.

7. REFERENCES

[1] S. Antani, D. Crandall, and R. Kasturi, "Robust extraction of text in video," In *Proc. 15th Intl. Conf. on Pattern Recognition*, Vol.1, pp. 831-834, 2000.

[2] T. Sato, T. Kanade, E.K. Hughes, and M.A. Smith, "Video OCR for digital news archive," In *Proc. IEEE Workshop on Content-Based Access of Image and Video Database*, pp. 52-60, 1998.

[3] X. Gao, and X. Tang, "Automatic news video caption extraction and recognition," In *Proc. 2nd Intl. Conf. On Intell. Data Engineering and Automated Learning*, 2000.

[4] D. Chen, K. Shearer, and H. Bourlard, "Text enhancement with asymmetric filter for video OCR," In *Proc. 11th Intl. Conf. on Image Analysis and Processing*, pp. 192-197, 2001.

[5] N. Otsu. "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man, and Cybernetics* ***(SMC-9)***, 9(1): 62-66, 1979.

[6] A. Rosenfeld. *Digital Picture Processing*. Academic Press. New York, USA, 1982.