A Chinese Character Localization Method based on Intergrating Structure and CC-Clustering for Advertising Images

Jie Liu  
Institute of Automation, Chinese Academy of Sciences  
Beijing, P.R. China  
jliu@hitic.ia.ac.cn

Shuwu Zhang, Heping Li, Wei Liang  
Institute of Automation, Chinese Academy of Sciences  
Beijing, P.R. China  
{swzhang, hpli, wliang}@hitic.ia.ac.cn

Abstract—In this paper, a novel Chinese character localization method is proposed for texts in advertising images. To deal with the texts with gradient color, a color clustering method based on edge is introduced to separate the color image into homogeneous color layers. To solve the problem of locating characters varied in size, style and arranged in irregular direction, a novel character localization method is proposed, which integrates structure and CC-clustering to locate characters according to reliable features of characters. Finally, a new noise removal method based on stroke width histogram is employed to remove all non-characters connected components, and then all characters are located. The experimental results show that the proposed method can effectively locate characters in advertising images.

Keywords—character localization; color clustering; connected component analysis

I. INTRODUCTION

Emerging techniques for ad monitoring and retrieval are of timely importance and interest. Text in advertising images is no doubt the most important clue for these purposes. Character localization is a fundamental step for performing these tasks.

Currently, there have been several studies concerned on character localization [1-10]. According to the features utilized, these methods can be broadly classified into two types: texture-based and region-based. Texture-based methods usually use texture analysis algorithms such as Gabor filtering [1], spatial variance [2] or wavelet transform [3] to locate text regions. Region-based methods use the properties of the color or gray scale in a text region or their differences with the corresponding properties of the background. Region-based methods can be further categorized into two sub-approaches: connected component (CC)-based and edge-based. CC-based methods [4-8] usually assume that text is represented with a uniform color. Therefore, they first quantize the color space of the input image into color layers by a clustering procedure, and then they analyze the connected components (CCs) and extract characters for each color layer. Edged-based methods [9-10] focus on the high contrast between the text and the background. Therefore, these methods identify the edges of text, and then filter out the non-text regions. Existing methods do solve the problem to a certain extent, however, not perfectly for text in ad images. The difficulty comes from the text variation in color, size, style and language. Besides, texts arranged in irregular direction also bring challenge to character localization.

In this paper, a new CC-based character locating method is proposed for ad images. Fig. 1 illustrates the framework of the proposed method. First color clustering method based on edge separates the color image into homogeneous color layers. Second the character localization method based on integrating structure and CC-clustering is performed to locate characters in every color layers. Finally a new strategy based on histogram of stroke width is introduced for noise removal. The major contributions of our approach are as follows:

1) Color clustering based on edge. There is no color clustering especially for characters. Classic color-clustering method cannot handle the text with gradient-color. Our method considers jointly two significant features of characters: similar color and sharp edges. Text with not only uniform color but also gradient color can be handled effectively.

2) The character localization method based on integrating structure and CC-clustering. In the advertising images, the aspects of characters are relatively stable while the noises vary irregularly. Based on this fact, the proposed method first extracts the features of the characters’ aspect, and then locates characters according to the features.

The rest of this paper is organized as follows: the color clustering method based on edge is described in section 2. In section 3 and 4, the proposed character localization method and noise removal method are presented. The detail of the experimental results are presented and discussed in section 5. Finally, we draw conclusions in section 6.

II. COLOR CLUSTERING BASED ON EDGE

The assumption of color homogeneity of a text is crucial to classical CC-based methods. However, it often does not hold in reality. There usually exists text with gradient color in ad images as shown in Fig. 2. There is no color clustering especially for characters. To solve the problem, we propose color clustering method based on edge.
hydis(l_i, l_j) = \text{dis}(l_i, l_j) \times \text{Max}(\log_{10} \text{dis}(p_i, p_j), \omega) \quad (1)

where \( i \) and \( j \) are indices of two adjacent pixels in an edge, \( l_i \) and \( l_j \) denote the colors of layers which contain pixel \( i \) and \( j \) in layer map respectively, \( p_i \) and \( p_j \) denote the colors of pixel \( i \) and \( j \) in original image respectively, \( \text{dis}(l_i, l_j) \) is the Euclidian distance between \( l_i \) and \( l_j \) in LUV color space, \( \text{dis}(p_i, p_j) \) is the Euclidian distance between \( p_i \) and \( p_j \) in LUV color space. The role of \( \text{Max}(\log_{10} \text{dis}(p_i, p_j), \omega) \) is to strengthen or reduce the effect of \( \text{dis}(l_i, l_j) \). When the value of \( \text{dis}(p_i, p_j) \) is small, the effect of \( \text{dis}(l_i, l_j) \) is reduced, and \( l_i \) and \( l_j \) are more likely equivalent colors, and vice versa. \( \omega \) is the threshold which prevents \( \text{dis}(p_i, p_j) \) from grossly reducing \( \text{dis}(l_i, l_j) \). It is set to 0.6 according to our experiences.

If the condition (2) is satisfied, \( l_i \) and \( l_j \) will be regarded as the equivalent colors.

\[ \text{hydis}(l_i, l_j) < T_i \quad (2) \]

where \( T_i \) denotes the threshold which is inversely proportional to the intensity of pixel \( i \) in edge map.

As we know, the edge may reflect the change from a component to another. The intensity of an edge reflects the extent of the change. The pixels with low intensity values in an edge may imply that they and their neighbors may belong to part of same component. Based on this fact, \( T_i \) is computed as

\[ T_i = \beta \times e^{-\varepsilon \log_{10}(I_i)} \quad (3) \]

where \( \beta \) and \( \varepsilon \) are coefficients, \( I_i \) denotes the intensity of pixel \( i \). \( T_i \) is inversely proportional to \( I_i \). We set \( \beta = 80 \), \( \varepsilon = 0.5 \) according to our experiences in our experiments.
The algorithm finds out the equivalent colors in the edges, and then merges the equivalent-color layers into one color layer.

III. CHARACTER LOCALIZATION BASED ON INTEGRATING STRUCTURE AND CC-CLUSTERING

The method consists of two stages: CC merging based on Chinese structure (CMCS), CC merging based on CC clustering (CMCC). CMCS is first used to generate sufficient candidate character, and then CMCC will extract the features of characters and locate characters according to the features.

A. CC Merging based on Chinese Character Structure

Based on some observations on characters in Chinese advertising images, some conclusions are drawn as follows:

- The ratio of aspect of normal Chinese character is approximately 1, and that of italic Chinese character is usually larger than 1 and less than 1.2.
- An interline spacing exists between parallel line of characters.
- A gap exists between characters in a text line.
- Almost all Chinese characters can be classified as one of the three character structuring patterns as shown in Fig. 5.

Based on these facts and three structures of Chinese characters, CC merging method based on Chinese character structure is proposed to generate sufficient candidate characters for next stage, CC merging based on CC-clustering.

Given two closely adjacent CCs, if the overlapping area of their circumscribing rectangles is greater than 60 percent of the area of smaller CC, they will belong to Inner-Outer structure. Otherwise the spatial relationship of the centers of their circumscribing rectangles will be used to determine them which belong to either Left-Right structure or Top-Bottom structure.

Reliable merging rule (RMR) and Weak reliable merging rule (WRMR) are important in this stage. Since RMR is more reliable than WRMR, it is encouraged in this stage. The overall process is demonstrated as Fig. 6.

RMR and WRMR are described as follow:

**RMR:**
If two closely adjacent CCs, CC$_i$ and CC$_j$, satisfy one of the following conditions, they will be merged into a whole CC.

1) CC$_i$ and CC$_j$ belong to Inner-outer pattern.
2) CC$_i$ and CC$_j$ belong to Top-bottom pattern and intersect.

Inner-outer pattern implies that two adjacent CCs could be parts of a character. Since a interline spacing exists between text lines, two intersected CCs belonging to Top-bottom pattern are mostly parts of a character.

**WRMR:**
If two closely adjacent CCs, CC$_i$ and CC$_j$, satisfy one of the following conditions, they will be merged into a whole CC.

1) CC$_i$ and CC$_j$ satisfy RMR.
2) CC$_i$ and CC$_j$ belong to Left-right pattern and they satisfy one of the following conditions:

- \[ \frac{W(CC_{\text{adj}})}{H(CC_{\text{adj}})} < k_1 \quad \text{and} \quad \frac{H(CC_{\text{adj}})}{W(CC_{\text{adj}})} < k_1 \quad \text{and} \quad \text{Dis}(CC_i, CC_j) < \min(L\text{Dis}, R\text{Dis}) \]
- \[ \frac{W(CC_{\text{adj}})}{H(CC_{\text{adj}})} < k_2 \quad \text{and} \quad \frac{H(CC_{\text{adj}})}{W(CC_{\text{adj}})} < k_2 \quad \text{and} \quad CC_i \cap CC_j \neq \emptyset \]

where CC$_{\text{adj}}$ denotes a new CC consisted of CC$_i$ and CC$_j$, W and H are width and height of the circumscribing rectangle of CC respectively, k1 can be set to 1.1 considering the fact that aspect ratio of normal Chinese character usually approximately is 1, similarly, k2 can be set at 1.2 since italic Chinese characters intersect horizontally and their aspect ratio usually is slightly larger than that of normal Chinese characters, Dis(CC$_i$, CC$_j$) denotes the distance between opposite sides of the circumscribing rectangles of CC$_i$ and CC$_j$, LDis denotes the distance between opposite sides of the circumscribing rectangles of the left adjacent CC and CC$_{\text{adj}}$, and RDis is defined in a similar way. To prevent CC$_i$ and CC$_j$ from incorrectly merging, Dis(CC$_i$, CC$_j$) must be less than the distances of their left and right adjacent CCs since a gap usually exists between characters.

B. CC Merging based on CC Clustering

Leader-follower clustering [12] is performed to cluster all CCs by the feature: width, height and aspect ratio. The features of cluster centers are regarded as reference features which are used to merge CCs.

The order of the reference features is crucial to this stage. The feature of a cluster containing many CCs is selected preferentially as reference feature according to
which merging CCs, and vice versa. Besides, the false combinations are unlikely caused by the reference feature with small aspect ratio. Based on these ideas, given a cluster $\text{Cluster}_i$, its priority $P(\text{Cluster}_i)$ is defined as

$$P(\text{Cluster}_i) = C(\text{Cluster}_i) \times R(\text{Cluster}_i)$$ (4)

where $C(\text{Cluster}_i)$ denotes the confidence term and $R(\text{Cluster}_i)$ denotes the reliability term, they are defined as

$$C(\text{Cluster}_i) = \frac{\text{Num}(\text{Cluster}_i)}{\sum_{j \neq i} \text{Num}(\text{Cluster}_j)}.$$ (5)

$$R(\text{Cluster}_i) = \frac{1}{A(\text{Cluster}_i)}.$$ (6)

Where $\text{Num}(\text{Cluster}_i)$ denotes the number of CCs in $\text{Cluster}_i$, $A(\text{Cluster}_i)$ denotes the width-height aspect ratio of $\text{Cluster}_i$, $n$ is the number of CC clusters.

CC merging rule by the reference features (7) is designed to merge as many adjacent CCs as possible.

$$W(\bigcup CC_i) \times \frac{1}{W(r_j)} < v_1$$ and $$H(\bigcup CC_i) \times \frac{1}{H(r_j)} < v_2$$ and $$\left| A(\bigcup CC_i) - A(r_j) \right| < v_3.$$ (7)

where $r_j$ denotes the reference feature $j$, $v_1, v_2$ and $v_3$ are thresholds. We set $v_1 = v_2 = 1.2, v_3 = 0.2$ according to our experiences in experiments.

For locating narrow or slim characters, such as ",", we need to do some special treatments. If a narrow CC cannot compose a character with other adjacent CCs while it satisfies the width of the reference feature, it will be accepted as narrow character.

The overall process can be demonstrated as as Fig. 7.

**Initialization:** Cluster CCs to generate the cluster set $C$; Compute priorities $P(\text{Cluster}_i), \text{Cluster}_i \in C$; Initialize the CC set $K$;

**Iterate**

Step 1: Choose and erase the $\text{Cluster}_i$ with the maximum priority from $C$; set the feature of $\text{Cluster}_i$ as the reference feature;

Step 2: Find as many adjacent CCs as possible which satisfy merging rule (8), merge them into a new CC and erase them from $K$;

**Until** The set $C$ or $K$ is empty

![Figure 7](image-url)
Figure 8. Some results of character localization by the proposed method.

<table>
<thead>
<tr>
<th>TABLE I. COMPARISON FOR CHARACTER LOCALIZATION</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>Recall rate</td>
</tr>
<tr>
<td>-------------------</td>
</tr>
<tr>
<td>Wang’s method</td>
</tr>
<tr>
<td>Our method</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TABLE II. COMPARISON FOR CHARACTER RECOGNITION</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recall rate</td>
</tr>
<tr>
<td>-------------------</td>
</tr>
<tr>
<td>Yang’s method</td>
</tr>
<tr>
<td>Our method</td>
</tr>
</tbody>
</table>

To evaluate the proposed method, we compare the performance of character localization with Wang’s method [6]. Recall and precision are the performance evaluation criterion. As shown in Table I, our method significantly outperforms Wang’s work for locating characters in the Chinese advertising images. The effectiveness of our approach is because the color clustering based on edge can correctly cluster the text with gradient color into one layer and the character localization method based on integrating structure and CC-clustering can locate characters according to more reliable features.

We also compare the proposed method with Yang’s method [13] by the same OCR software. The recognition rate is criterion of evaluating the performance. Table II presents the comparison results. The results show that our method achieves higher recognition rates. This is due to the capability of handling the text varied in size and style and the text arranged in irregular direction. Yang’s method cannot perfectly deal with these cases; therefore it generates relatively lower recognition rates.

VI. CONCLUSION

In this paper, a novel method is proposed to locate character in ad images, which is composed of color clustering method based on edge, character localization method based on integrating structure and CC-clustering and noise removal method based on histogram of stroke width.

By using this color clustering method, the text with gradient color can be correctly dealt with. By integrating structure and CC-clustering, the features of characters can be automatically extracted so that the characters can be efficiently located according to these reliable features. The experimental results show that the proposed method is robust for the variation of character in color, size and style. Besides, it can also handle text arranged in irregular direction.

ACKNOWLEDGMENT

This work has been supported by the National Key Technology R&D Program of China under Grant No. 2009BAH48B02, 2009BAH43B04, 2011BAH16B01 and 2011BAH16B02. The authors thank the anonymous reviewers for valuable comments.

REFERENCES