Newspaper Headline Extraction Method Based on Reversed Region Geometric Analysis[*]

Guowang Miao, Liangrui Peng, Xiaoqing Ding

State Key Laboratory of Intelligent Technology and Systems

Dept. of Electronic Engineering, Tsinghua Univ., Beijing, 100084, P. R. China

Abstract:

The OCR(Optical Character Recognition) system will not be able to recognise the characters in the newspaper headlines on graphical designs with partially reversed regions. This paper gives a method to extract the character string based on the geometrical characteristics analysis of the headline's reversed region boundary. Experiment results show this method is effective.

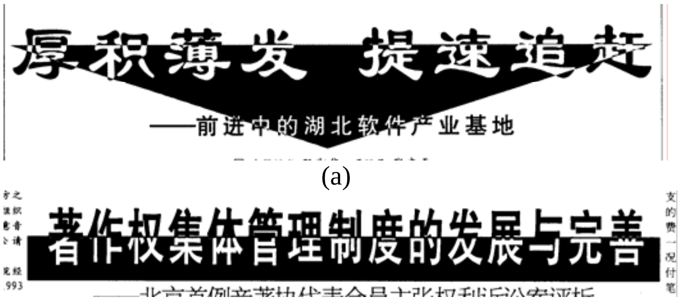
Keywords: *character string extraction, image preprocessing, newspaper headline*

1. Introduction

At present, the OCR (Optical Character Recognition) system has already reached a rather high recognition rate for text characters, but most OCR systems cannot recognise these characters in newspaper headlines with special background design, as shown by Fig.1. Because the systems target traditional character strings like black characters on a plane white background. In order to recognise these characters successfully, preprocessing is required to convert these characters into traditional type before recognition. Existing preprocessing techniques, however, could not extract the character strings from the reversed regions with graphical design.

Takebe et al.[1] tried to extract the characters from images with a background design. His method needs that the character to be extracted have the monolithic colour, and also, it must iteratively calculate the resemblance value of the characters during preprocessing, which has a rather low efficiency. This method could not handle the situations showed in Fig.1, which the characters to be extracted are partially reversed.

This paper presents an effective method of extracting the character string from the newspaper headlines with reversed graphical background. We assume that the headlines regions have been located by layout analysis module before our procedure.



(a)

(b)

Fig. 1: Examples of headlines with graphical designs in Chinese newspapers

In most headlines with reversed regions, the reversed regions are in geometrical shapes. The reversed geometric region usually has the longest borderline. We first extract all the black-and-white edge of the headline image, and search out the longest black-white edge. After analysing the characteristic points and the characteristic lines of the extracted edges, we can determine the shapes of the reversed region. Based on some rules, we correct the edges into a whole continuous closed boundary, which exactly encloses the region that need to be reversed. Then what we need to do is to reverse all pixels that are enclosed by the boundary. We can finally get an ideal result we've expected.

Section 2 describes the method and system implementation in detail. Section 3 shows experiment results. Conclusion is given in Section 4.

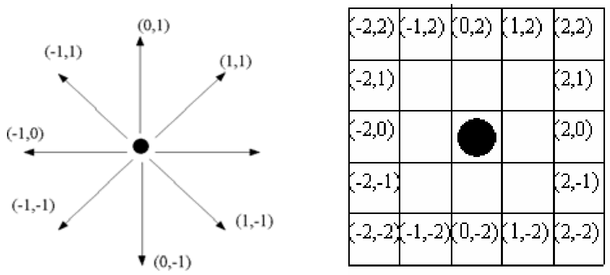
2. Character String Extraction

2.1 Boundary Extraction

In order to find the longest boundary of the reversed region of the newspaper headline image, we first search N(in our project N=20) starting points at the interface of black and white pixels which distribute uniformly in the

[*] Supported by Tsinghua SRT Plan (project 11125) & National Natural Science Foundation of China (project 69972024)

headline. Then we search out the black-white boundary deriving from each starting point. The way we search the next black-white boundary points is based on the 8-connectivity of the segments, where the directions of each segment is coded using a coordinate scheme shown in Fig. 2 (a). If we failed to find a boundary point in its 8-connectivity neighbourhood due to some noise, enlarge the searching circle to the 16-connectivity of the segments, whose direction coordinate scheme is shown in Fig. 2 (b). In this way, we can take out the most accurate boundary that maybe unclosed, but has already contained enough information which will bring out good processing effect in the following steps. Every searched boundary point is marked so as not to be searched twice.



(-2,2)	(-1,2)	(0,2)	(1,2)	(2,2)
(-2,1)				(2,1)
(-2,0)		●		(2,0)
(-2,-1)				(2,-1)
(-2,-2)	(-1,-2)	(0,-2)	(1,-2)	(2,-2)

(a)

(b)

Fig. 2 Directions for (a)4-directional code, and (b) 16-directional code.

After calculating the number of points in each boundary, we get the one comprising the most boundary points and save all the positions of its elements. Fig. 3 shows the longest boundary extracted from headline image in Fig. 1(a).

2.2 Characteristics Extraction and Shape Determination

After the observation of lots of extracted longest borderlines, we get some conclusions that fit most cases of the headlines which consist of linear-boundary reversed region.

In headlines on graphical designs, the shape of the reversed regions frequently used is triangle or rectangle, or the combination of them. The reversed region is usually upright. In other words, the rectangular boundary is either vertical or horizontal, and the triangle has at least one vertical or horizontal edge.

Usually, most extracted points are located on the triangle or rectangle border compared with the other lines. In some special situations, not all the vertices of the reversed region are contained in the extracted boundary points. But they’re located at the crossing points of two extended lines which may be extracted by the boundary points we’ve got.

Based on the analysis above, we extract the characteristics of the region in the following way with the flowchart shown in Fig. 4.

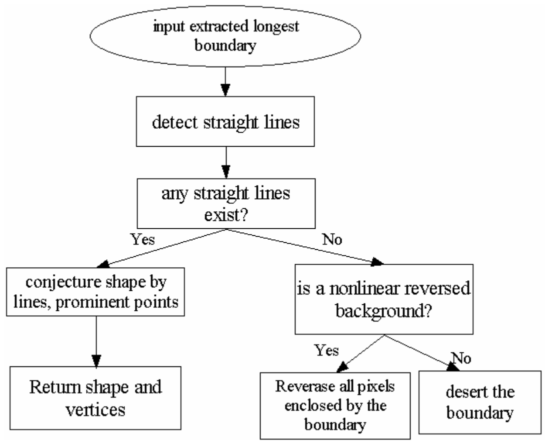

input extracted longest boundary
detect straight lines
any straight lines exist?
Yes
No
conjecture shape by lines, prominent points
is a nonlinear reversed background?
Yes
No
Return shape and vertices
Reverase all pixels enclosed by the boundary
desert the boundary

Fig. 4 Flowchart of determining the reversed region shape

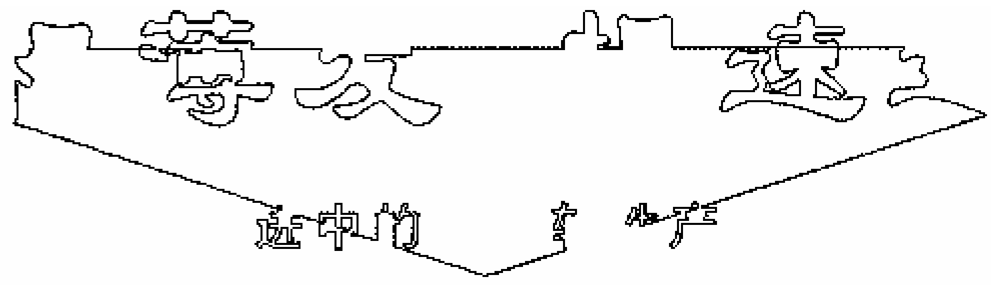
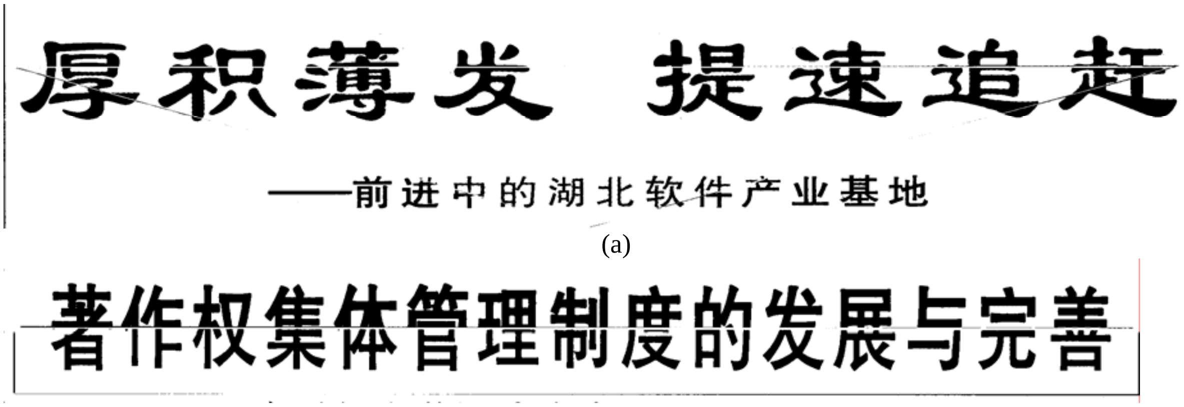


Fig. 3 Sample of boundary extracted from headline with graphical design(extracted from Fig. 1(a))

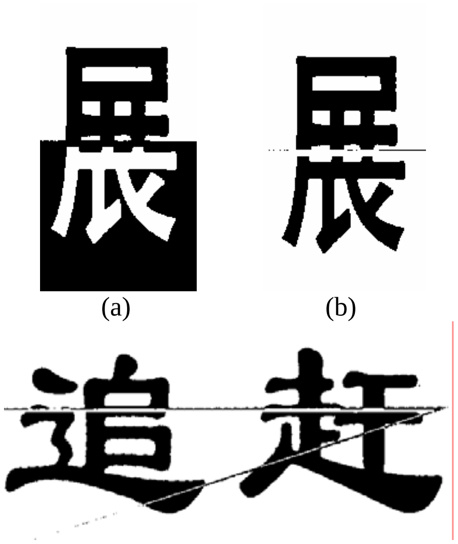


(a)

(b)

Fig. 5 Samples of headlines preprocessed with regular region reversed ((a)from Fig. 1(a);(b) from Fig1.b

We try to find the points in the boundary points that are prominent. And based on these points, get the straight lines contained in the boundary. If no straight lines are found, judge whether there is the nonlinear reversed region.



(a) (b)

(c)

Fig. 6 Samples of failed-reversed region a the original character ;b,c the character not ideally processed

Here we just reverse the approximate region we found. In Fig. 5, we show some images that have been processed.

As we see in Fig. 5, most regions have been reversed correctly. but the place located at the edge cannot be processed ideally (See Fig. 6 in detail). There are two main reasons. First, the boundary of the region we reverse is ideal. In the real printings, the printing ink always go through the boundary from place to place, which make quite a lot of noises. Second, the boundary we've determined does not exactly match the ideal line (See Fig. 6 (c)). We call such phenomenon “edge effect”. The edge effect needs further processing to get a better result for OCR system.

2.3 Boundary Correction

When edge effect exists, the boundary extracted includes points of both the black background and black characters. Obviously, the boundary points of characters are redundant. Hence, we'll delete the redundant points and fill new points into the boundary to complement it to be a closed one that precisely encloses the region to be reversed.

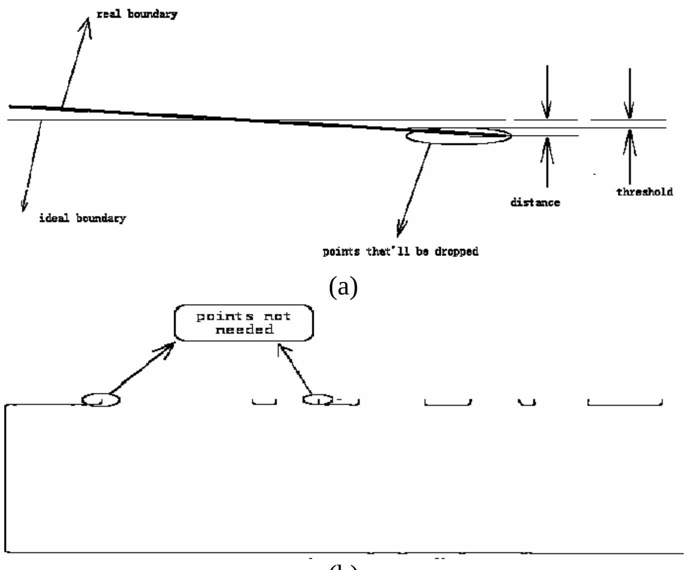
i) Redundant Boundary Points Deletion

Here we take two steps to delete the redundant boundary points. The approximate region and its boundary, which we call ideal boundary below, have already been determined. Thus any original boundary points whose perpendicular distance from the determined boundary exceeds a preset threshold is deleted as redundant boundary points. The threshold should neither be too small nor too big, it should be compatible with the average stroke width of the character image.

One shortcoming to mention is that if the headline is a little bit skew even after the de-skewing process, as is shown in Fig. 7(a), the distances of the points located at the two ends of the straight line may exceed the threshold, which makes these points deleted.

After the operation of distance discrimination, most redundant points have been deleted. And still there are some points that we do not want which are too close to the ideal boundary. But they have a common ground that they are all prominent compared with the points next to them, shown in Fig. 7 (b). With this characteristic, we can drop these points by the discrimination of slope.

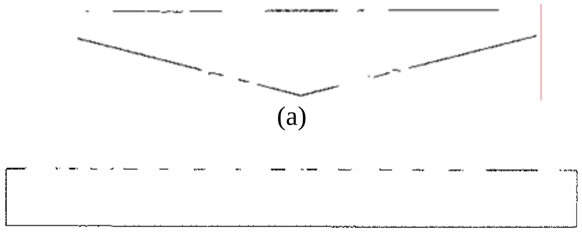
Here we show two boundaries that have few redundant points in Fig. 8.



(a)

(b)

Fig. 7Example of points mistakenly dropped



(a)

(b)

Fig. 8 Original boundary without redundant points

ii) Fragmentary Edges Linking

We then need to complement those fragmentary edges into a closed continuous boundary. If the boundary is broken somewhere, we insert points between the two break points. The points inserted satisfy the following two conditions (assume the straight line that goes through the two break points is A):

1) The inserted point is a interconnected point of black and white spaces, whose perpendicular distance from line A does not exceed a present threshold.

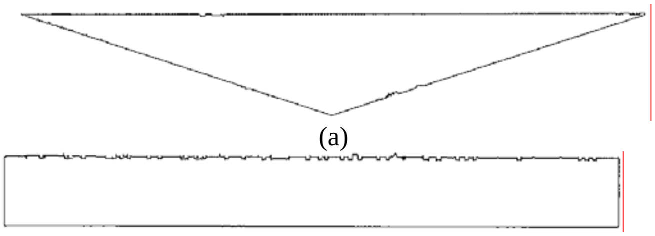
2) If no points are found that satisfy condition 1, just insert points between the two break points linearly.

After this operation, we'll get the optimised boundary. Here're two samples (Fig. 9).

We've already get a precise boundary of the region which need reversing. Hence, what we have to do next is just to reverse all pixels enclosed by the boundary. Some headlines after preprocessing are shown here (Fig. 10). From the samples shown in Fig. 10,we can see that the edge effect has been overcame ideally.

2.4 System Implementation

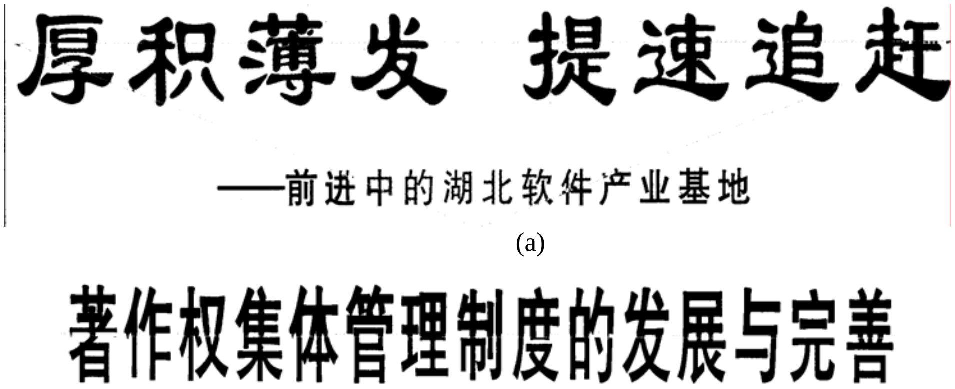
Based on the general ideas above, we realise our system, whose flowchart is shown in Fig.11.



(a)

(b)

Fig. 9 Samples of optimized boundary



(a)

(b)

Fig. 10 Headlines processed without "edge effect"

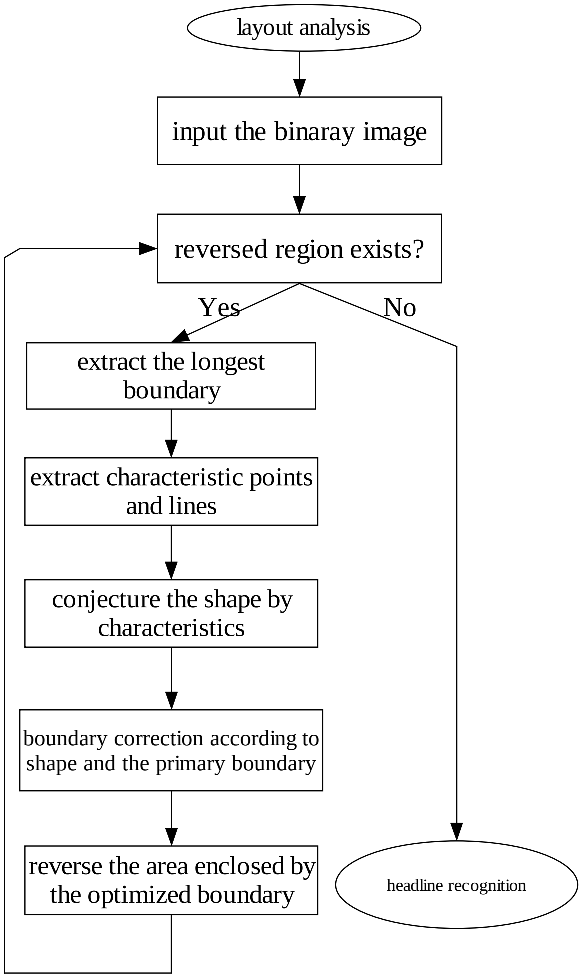

layout analysis
input the binaray image
reversed region exists?
Yes
No
extract the longest boundary
extract characteristic points and lines
conjecture the shape by characteristics
boundary correction according to shape and the primary boundary
reverse the area enclosed by the optimized boundary
headline recognition

Fig. 11 Flowchart of the preprocessing operation

When we get a headline, judge whether there're reversed regions. The judgement is done by calculating the portion of black pixel's density to white one's. If the portion is bigger than a threshold, we decide that there exists a reserved region. Then we extract the longest boundary with coordinates saved. After analysing the characteristic points and lines, the shape of the reserved region is determined. In order to reverse precisely the region we want, some boundary corrections are done. So we get the exact region that need to be reversed, and finally reverse all pixels enclosed in the optimised boundary.

3 Experiment and Result

We applied on this method to 106 headlines of different types of reserved region. And 2 headlines failed to be correctly reversed. The correct extraction rate was about 98.11%. The unit applied to this extraction rate is a headline. We consider an extracted string with more than one character mistakenly reversed to be an error. In Fig. 12 and Fig.13,we show some examples of the experiment.

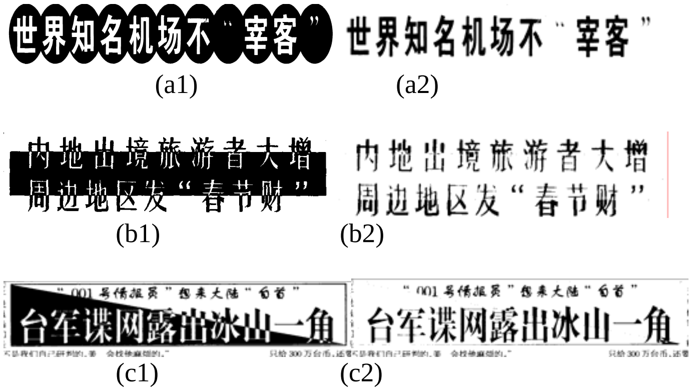


Fig. 12 Examples of successefully reversed headlines

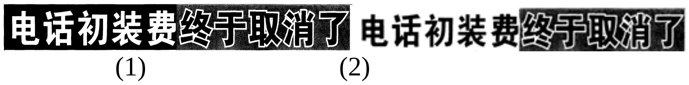


Fig. 13 Examples of miatakenly processed headlines

Fig. 13 shows an example with mixed types. Half of the title is reversed, which can be processed by the method presented in this paper. The other half part is with dark textural background, which should be handled by other method.

4 Conclusion

This paper proposed a method for preprocessing headlines with fully or partially reversed graphical designs. We reverse the regions according to the geometrical characteristics of their boundary, and in order to get a better effect, we deal with the "edge effect" specially. Experimental results indicate that our method is effective.

Future research will focus on preprocessing method for other types of newspaper headline designs, such as headline with textural background, hollow characters, etc.

References

[1] Hiroaki Takebe, Yutaka Katsuyama, and Satoshi Naoi: "Character string extraction from newspaper headlines with a background design by recognizing a combination of connected components" Part of the IS&T/SPIE Conference on Document Recognition and Retrieval VI, San Jose California, January 1999

[2] Satoshi NAOI, Maki YABUKI, Atsuko ASAKAWA, Yoshinobu HOTTA "Global Interpolation in the Segmentation of Handwritten Characters Overlapping a Border" . IEICE TRANS. INF. &SYST. VOL. E78-D.NO-7 JULY 1995