

# AN APPROACH TO WORD IMAGE MATCHING BASED ON WEIGHTED HAUSDORFF DISTANCE

*Yue Lu, Chew Lim Tan, Weihua Huang, Liying Fan*

Dept. of Computer Science, National University of Singapore, Singapore 119260

## ABSTRACT

*An approach to word image matching based on weighted Hausdorff distance(WHD) is proposed in this paper to facilitate the detection and location of the user-specified words in the document images. Preprocessing such as eliminating the space between adjacent characters in the word images and scale normalization is first done before the WHD is utilized to measure the distance between the template image and the word image extracted from the document image. Experimental results in the application of detecting the user-specified words from both English and Chinese document images show that it is a promising approach for word image matching.*

## 1. INTRODUCTION

Many commercial Optical Character Recognition(OCR) systems, which claim to achieve the recognition rate of above 99% for single characters, have been introduced to the market during the past few decades. However, when confronted with degraded images, the performance of these systems deteriorates severely because all of them rely on the segmentation procedure that is prone to recognition failure at the presence of image noise and poor printing quality. In particular, when adjacent characters are joined or fused, an OCR system must perform the delicate task of separating them. However, under some conditions perfect segmentation would be impossible. It has been estimated that half of the errors in the character recognition occur during segmentation in locating the individual characters within a word.

To overcome the problem caused by character segmentation, segmentation-free approaches have been developed by eliminating the segmentation step. Such approaches have been reported by using Hidden Markov Model and Viterbi dynamic programming algorithm in handwritten word recognition<sup>[1]</sup>. These strategies can be viewed as searching for the most likely sequence of characters given a sequence of observations extracted from the input image. A lexicon can be an available tool for

resolving the ambiguity during the procedure. Another approach is called holistic word matching<sup>[2]</sup>, which treats the word as a single, indivisible entity and attempts to recognize it using features of the word as whole, rather than segments and recognizes individual characters of the word. The latter approach is inspired by psychological studies of human reading which indicated that humans use features of word shape such as length, ascenders, and descenders in reading.

Word matching may be at either the feature-level or the pixel-level. As a low-level matching, the pixel-level matching, such as Hausdorff distance, is simple and insensitive to changes of image characteristics. Hausdorff distance measure has been widely investigated in the area of object matching<sup>[3-5]</sup> and character recognition<sup>[6]</sup>. Dubuisson and Jain<sup>[4]</sup> presented the modified Hausdorff distance(MHD) based on the average distance value for the object matching. For the purpose of word image matching, we propose a weighted Hausdorff distance(WHD) in this paper to improve the MHD by setting the different parts of the word image with different contributions to the directed distance.

The remainder of this paper is organized as follows. Section 2 discusses the preprocessing for the word image. Section 3 presents the weighted Hausdorff distance measure. Section 4 describes the experimental results that demonstrate the effect of the proposed approach in detecting and locating user specified words in English documents and Chinese documents. Finally, conclusion is given in Section 5.

## 2. PREPROCESSING FOR WORD IMAGES

The preprocessing for word image matching includes space elimination and scale normalization.

In a word image, it is common that two or more adjacent characters are connected with each other, which is possibly caused by low scanning resolution or poor printing quality. Generally speaking, it is not trivial to separate them.

On the other hand, the templates of word images used for the matching are synthesized directly from bitmap

images of each character one after another, in which each character occupies a uniform size of image pixels, e.g. 16\*16 pixels for each one. This results in variable spacing between adjacent characters. From figure 1(a), we can find that the spacing between the double characters “l” is larger than that between the character “w” and “a”.

To overcome the problems, we condense the characters in the word image by eliminating all of the space between each part of the adjacent characters in both the template image and the word image extracted from the document image, as showed in Figure 1(b) and Figure 2(b).

And then, the word image is normalized to the size of template image.

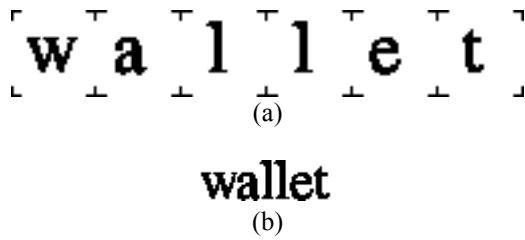


Fig. 1 word template

(a) Original template image (b) Condensed image

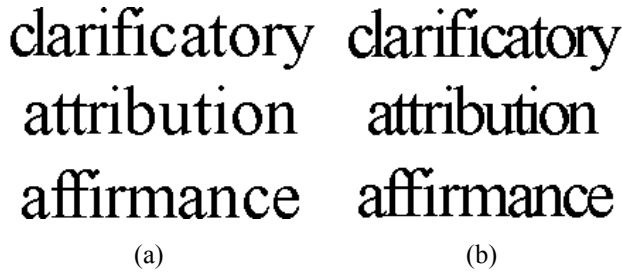


Fig. 2 Space elimination

(a) Original word images (b) Space eliminated images

### 3. WORD IMAGE MATCHING BASED ON WEIGHTED HAUSDORFF DISTANCE

Hausdorff distance has been widely applied in two-dimensional image matching, especially in the area of object matching.

The distance(e.g. Euclidean distance) between two points  $a$  and  $b$  is defined as  $d(a,b)=\|a-b\|$ , and the distance between a point  $a$  and a finite point set  $B=\{b_1,\dots,b_{N_b}\}$  is commonly defined as

$$d(a,B)=\min_{b \in B} \| a - b \|$$

Given two finite point sets  $A=\{a_1,\dots,a_{N_a}\}$  and  $B=\{b_1,\dots,b_{N_b}\}$ , the Hausdorff distance is defined as:

$$H(A,B)=\max(h(A,B),h(B,A))$$

Where  $h(A,B)$  and  $h(B,A)$  represent the directed distance between two sets  $A$  and  $B$ . The directed distance  $h(A,B)$  is traditionally defined as

$$\begin{aligned} h(A,B) &= \max_{a \in A} d(a, B) \\ &= \max_{a \in A} \min_{b \in B} d(a, b) \\ &= \max_{a \in A} \min_{b \in B} \| a - b \| \end{aligned}$$

The function  $h(A,B)$  identifies the point  $a \in A$  that is farthest from any point of  $B$  and measures the distance from  $a$  to its nearest neighbor in  $B$ . The Hausdorff distance  $H(A,B)$  measures the degree of mismatch between two point sets  $A$  and  $B$ .

Dubousson and Jain<sup>[4]</sup> presented the modified Hausdorff distance(MHD) measure by employing the summation operator over all distance, rather than the maximum operator:

$$h_{MHD}(A,B)=\frac{1}{N_a} \sum_{a \in A} d(a, B)$$

The MHD was proposed for the purpose of object matching in the areas of computer vision, object recognition and image analysis.

A word image can be divided into different parts, namely the ascender, the descender, and the mid zone, as illustrated in Figure 3. We propose a weighted Hausdorff distance(WHD) to investigate the application of Hausdorff distance to word image matching, in which the contribution of different parts of the word image to the Hausdorff distance is not the same. The directed distance of WHD is computed as

$$h_{WHD}(A,B)=\frac{1}{N_a} \sum_{a \in A} w(a) \cdot d(a, B)$$

where

$$\sum_{a \in A} w(a) = N_a$$

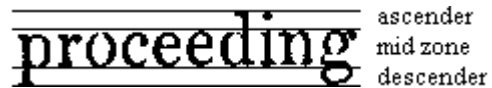


Fig. 3 different parts in the word image

Figure 4 illustrates five word images extracted from real document images. To prove the effect of the proposed WHD, the distance measures between the five word images and their template images that are generated by the bitmap images of the characters are calculated.

The preprocessing is carried out for both the word images and template images using the method described in Section 2 before the matching. The weight of the different parts in the word namely  $w(a)$ ,  $w(d)$  and  $w(m)$  correspond

to the ascender, the descender and the mid-zone respectively. They are set as:

$$w(a)=w(d)=2*w(m)$$

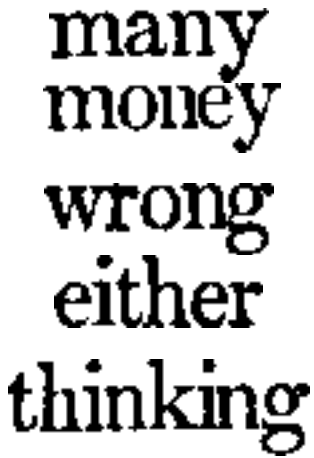


Fig. 4 word images extracted from real documents

As comparison, the distance measures of both MHD and WHD are reported in Table 1-5. The images 1-5 in the tables correspond to the five images illustrated in Figure 4.

Table. 1 distance with template image of word “many”

	Image1	Image2	Image3	Image4	Image5
MHD	0.211	0.338	0.485	0.834	0.595
WHD	0.155	0.350	0.491	0.762	0.598

Table. 2 distance with template image of word “money”

	Image1	Image2	Image3	Image4	Image5
MHD	0.448	0.269	0.482	0.868	0.538
WHD	0.452	0.261	0.579	0.841	0.550

Table. 3 distance with template image of word “wrong”

	Image1	Image2	Image3	Image4	Image5
MHD	0.480	0.407	0.191	0.903	0.580
WHD	0.492	0.417	0.187	0.998	0.550

Table. 4 distance with template image of word “either”

	Image1	Image2	Image3	Image4	Image5
MHD	0.727	0.743	0.935	0.187	0.571
WHD	0.827	0.796	0.969	0.165	0.589

Table. 5 distance with word template image of “thinking”

	Image1	Image2	Image3	Image4	Image5
MHD	0.757	0.648	0.705	0.942	0.226
WHD	0.778	0.685	0.738	0.976	0.196

From the results, we can see that the distance measure represented by WHD between each word image and its corresponding template image is lower than that represented by MHD, whereas the distances of WHD between the template image and other images are larger

than that of MHD generally. It means that WHD is better than MHD for the purpose of word image watching.

#### 4. EXPERIMENTAL RESULTS

To verify the validity of the approach of word image matching based on the proposed weighted Hausdorff distance(WHD), we use it to detect and locate the user-specified words in the document images. The experiments are carried out on both English document images and Chinese document images.

The document images are selected from the scanned books and newspapers that are provided by the Central Library of the National University of Singapore. Before the word image matching, all connected components in the document images are extracted, and the topological position relations among the connected components are utilized to bound the word images.

##### 4.1 English document image

Firstly, the space between adjacent characters in both the template image of specified word and the images of words extracted from the document is removed, then the word images are normalized to the size of the template image before the word matching. Some words are eliminated simply by their shape ratio. Suppose that  $R_w$  and  $R_t$  represent the shape ratio of word image and template image respectively. Only the word images whose  $R_w/R_t$  values fall in the range of  $(\lambda_1, \lambda_2)$  will be matched with the template image. In our experiments,  $\lambda_1$  and  $\lambda_2$  are set as 0.8 and 1.2 respectively.

Figure 5 demonstrates an example, in which the user-specified word “military” is detected and located successfully in the document image by within the bounding rectangles.

Even the soldiers are forced to become agriculturists. In certain parts of the empire, fields are laid out, which they cultivate for their subsistence; in other parts, where they have no farms of their own, they hire themselves out as servants to the peasants, and plough the fields, till they are called for the military reviews. The greater part of the officers are very illiterate, and have risen from the ranks. There are, however, military examinations, as well as literary, and degrees of bachelor, master, and doctor, in military tactics, regularly conferred. They have the same degrees as literary mandarins, and wear the same badges of rank, buttons or knobs, on their caps; yet they are regarded both by the literary mandarins and the people with the greatest contempt. Their salary is very small, their resources slender, and their situation not at all enviable. Many of the general officers are Tartars, who enjoy great

Fig.5 Detection of specified word in an English document

## 4.2 Chinese document image

Chinese words are quite different from the western words. It is reported by psychological research that the four corners are more important in the recognition of characters for human beings. We divide a Chinese character image into different parts, as showed in Figure 6. The weight for the WHD is variable according to different parts. We set

$$w(c)=2*w(h)=3*w(p)$$

where  $w(c)$ ,  $w(h)$  and  $w(p)$  represent the weight for the corner, heart and peripheral part respectively.

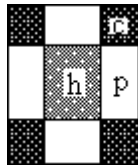


Fig. 6 Different parts in a Chinese character

(吉隆坡讯) 马来西亚外交部长赛哈密否认新加坡的谈话，指巫统选举已影响到两国针对一系列悬而未决的问题的谈判。

他指出，由于新加坡对一些课题拒绝承诺，使到谈判陷入僵局。

他是回应新加坡外交部长贾古玛的谈话发表这项评论。据新加坡《海峡时报》3月9日引述贾古玛的谈话说，由于忙于应付巫统党选，马国政府并未对下一轮的会谈日期作出反应。

赛哈密认为，这种看法应该纠正过来，事实上是“新加坡政府并未对（我们）向他们提出的建议作出反应。”

Fig.7 Detection of specified word in a Chinese document

In General, one Chinese word/phrase is composed of several successive characters. We calculate the distance of each corresponding character of the word, and take the minimum of them as the final measure.

Figure 7 gives the result in which the Chinese word “Singapore” is detected and located successfully from the document image.

## 5. CONCLUSIONS

We proposed an approach of word image matching based on weighted Hausdorff distance in this paper. The performance of the weighted Hausdorff distance is better than that of modified Hausdorff distance for the purpose of the word image matching. The experimental results in the application of detecting the user-specified words from document images show that the approach is promising for word image matching.

**Acknowledgements** The authors would like to thank Mr. Ng Kok Koon and Mrs. Khoo Yee Hoon of the Central Library of the National University of Singapore for providing us the document images. This project is supported by the National Science & Technology Board and Ministry of Education under research grant R255-000-071-112/303.

## REFERENCES

- [1] A. Kundu, “handwritten word recognition using Hidden Markov Model,” H. Bunke and P. S. P. Wang(Ed.), Handbook of character recognition and document image analysis, World Scientific, 1997, pp.157-182
- [2] S. Madhvanath, E. Kleinberg and V. Govindarju, “Holistic verification of handwritten phrases”, IEEE trans. Pattern Analysis and Machine Intelligence, Vol.21, No.12, pp1344-1356, 1999.
- [3] D. P. Huttenlocher, G.A. Klanderman, and W. J. Rucklidge, “Comparing images using the Hausdorff distance”, IEEE trans. Pattern Analysis and Machine Intelligence, Vol.15, No.9, pp850-863, 1993.
- [4] M. P. Dubuisson and A. K. Jain. “A modified Hausdorff distance for object matching”, in Proc. 12th Int. Conf. Pattern Recognition, Jerusalem, Israel, Oct. 1994, pp.566-568
- [5] D. G. Sim, O. K. Kwon, and R. H. Park, “Object matching algorithm using robust Hausdorff distance measures”, IEEE trans. Image Processing, Vol.8, No.3, pp425-425, 1999.
- [6] X. Wu, P. Shi, Unconstrained handwritten numeral recognition using Hausdorff distance and multi-layer neural network classifier. Proceedings of the Fifth International Conference on Document Analysis and Recognition, 1999. pp: 249 –252