WORDS RETRIEVAL FROM TEXT IMAGES

Nikolay Kirov Kirov
Computer Science Department, NBU and
Institute of Mathematics and Informatics, BAS
e-mail: nkirov@nbu.bg

ABSTRACT

In this paper we present results of applying Hausdorff type distance for searching words in a set of graphic files representing pages of scanned book. For successive retrieval, a number of parameters are used. We investigate the influence of image resolution and point distance over searching results.

1. INTRODUCTION

The main question in this paper is: How to find a word in a text document? When the document is represented as text file, the answer is quite trivial - open the file in any text editor, choose a word and push Find button. But the task is not so easy when the document is a set of graphic images. This is natural situation when we deal with digitization of cultural and scientific heritage and when scanner devices produce files in graphic formats.

Optical character recognition (OCR) is the usual way of conducting text retrieval from scanned document images. OCR software converts text images into a text file, recognizing every letter and mapping it to a number, which is called code. This technique is well developed and has high accuracy. And then we apply the previous algorithm. But sometimes OCR is very difficult process requiring dictionaries in the corresponding language. Often human efforts are needed to correct OCR errors. Here are some obstacles to successful OCR:

- The quality of page images;
- Language dependency (alphabet and coding, unknown language):
  - dictionaries;
  - old grammar, obsolete words and phrases and idioms;
  - old letters, out of the coding tables;
  - multi-lingual documents;
- Errors in automatic OCR, human intervention needed.

We suggest a different approach: instead of applying two steps - OCR and searching in text documents, we try searching words directly in a scanned text documents. We can organize retrieval of words, similar to a given pattern word, (searching in the binary text images). Similar ideas can be found in [5] and [7].

Three main steps are essential for successful word searching: segmentation, searching and result representation. In the segmentation step we create so-called word images - every word is encompassed by a rectangle, which consist of white and black

This research has been partially supported by a Marie Curie Fellowship of the European Community program ``Knowledge Transfer for Digitalization of Cultural and Scientific Heritage in Bulgaria" under contract number MTKD-CT-2004-509754.

pixels. For measuring similarities between word images we use Hausdorff type distances (see [3]). Choosing a concrete Hausdorff distance, we have freedom to use various point distances. In this paper we consider some distances (on the plane) and compare the results of searching a word in a set of scanned pages of a book.

2. SEGMENTATION

We use horizontal projection for row extraction (see Fig. 1). If the rows are horizontal (straight lines), the histogram has near zero values between rows. The same case is when the rows have small slopes.

Vertical projection is a common method in character or word segmentation. The histogram is obtained by counting the number of black pixels in each vertical scan at a given horizontal position (see Fig. 2). If the characters are well separated, the histogram should have zero values between the characters. Because the distances between words are larger than between characters, it is easier to separate words than characters.

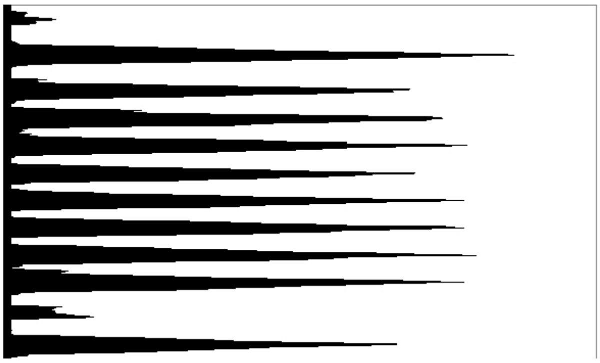


Figure 1: The horizontal histogram of a page

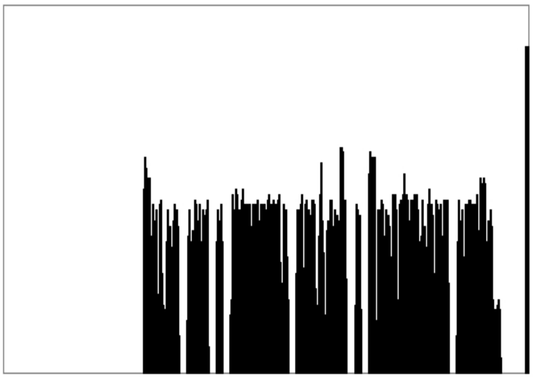


Figure 2: The vertical histogram of a row

For segmentation step we use a number of parameters p_1, p_2, p_3, p_4, which are important for successful word separation.

- The height of every row must be at least p_1. This helps us to avoid creating (due to noise) rows with small height;
- When the value at a point in row histogram is less than p_2, we suppose that this point belongs to a white space between the words.
- The white space between words must be greater than p_3. This helps us to separate word images from some special symbols as dots, commas, etc.
- The parameter p_4 concerns additional step conducted when we have already separated words, and word images are framed. At this step we try to shrink the frame rectangles from top and bottom. We use horizontal and vertical histograms only for the points in a given word image. We decrease the height of rectangle if the points of horizontal histogram have values less than p_4. This step is very useful when the rows have small slopes (see Fig. 3).

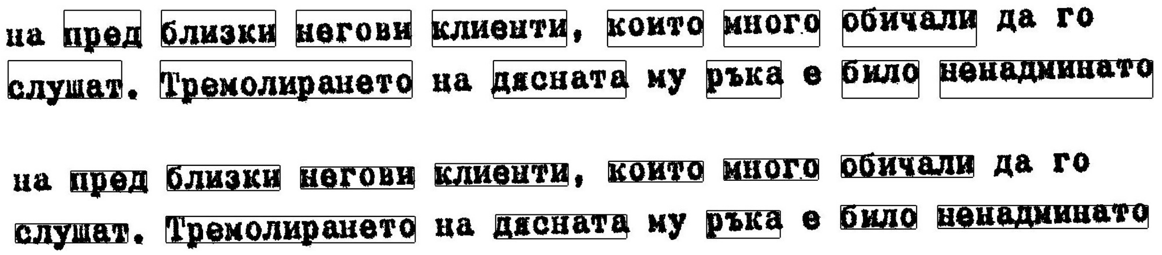

на пред близки негови клиенти, които много обичали да го слушат. Тремолирането на дясната му ръка е било ненадминато
на пред близки негови клиенти, които много обичали да го слушат. Тремолирането на дясната му ръка е било ненадминато

Figure 3: Small slope of rows: word segmentation before and after the step “shrink”

3. SEARCH

3.1 HAUSDORFF TYPE DISTANCES

The Hausdorff type distances between the sets of points on the plane are commonly used as similarity measures for binary images. The classical Hausdorff distance (HD) between two point sets A and B is defined as

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

where $h(A,B)$ and $h(B,A)$ are so-called directed distances between the sets. For original Hausdorff metrics $h(A,B) = \max_{a\in A} d(a,B)$, where $d(a,B) = \min_{b\in B} \rho(a,b)$ is the distance from a point a to the set B. $\rho(a,b)$ is any point distance.

For image matching a number of modifications of $h(A,B)$ have been introduced by many authors. Dubuisson and Jain [4] consider so-called Modified Hausdorff Distance (MHD), one of the best methods for word search in the text images (see also [3]). They replaced $h(A,B)$ by

$$h_{\mathrm{M}HD}(A,B) = \frac{1}{N_A}\sum_{a\in A} d(a,B) = \frac{1}{N_A}\sum_{a\in A}\min_{b\in B}\rho(a,b),$$

where N_A is the number of points in set A. A bit better results were obtained in our examples omitting the coefficient $\frac{1}{N_A}$ in front of the sum (2). We called this modification Sum Hausdorff Distance (SHD), [2]

$$h_{SHD}(A,B) = \sum_{a\in A} d(a,B) = \sum_{a\in A}\min_{b\in B}\rho(a,b).$$

The directed distance $h_{\mathrm{M}}(A,B)$ for M-HD [6] is defined by

$$h_{\mathrm{M}}(A,B) = \frac{1}{N_A}\sum_{a\in A} f(d(a,B)),$$

where the function f is

$$f(x) = \begin{cases} |x| & \text{if } |x| \le \tau \\ \tau & \text{if } |x| > \tau. \end{cases}$$

This means that we sum the distances $d(a,B)$ which are less than the constant τ and add τ when the distance is greater than τ. The authors of [6] recommended $\tau \in [3,5]$. Note that MHD with any bounded point distance is M-HD.

Detailed HD distance measures comparisons can be found in [3]. We use M-HD in our experiments with $\tau = 5$ because this simplifies the computations and speed up the searching process.

3.2 POINT DISTANCES

Let $a = (a_x, a_y)$ and $b = (b_x, b_y)$ are two points on the plane. Well known Euclidean distance is

$$\rho_2(a,b) = \sqrt{(a_x - b_x)^2 + (a_y - b_y)^2}$$

also called Minkowski distance of order 2. Manhattan distance or Minkowski distance of order 1 is

$$\rho_1(a,b) = | a_x - b_x | + | a_y - b_y | .$$

The infinity norm distance

$$\rho_{max}(a,b) = \max\{| a_x - b_x |, | a_y - b_y |\}$$

is often used in the applications too. The last two variants are easy to be calculated, without multiplication and using square root. Because $\rho_{max}(a,b) \le \rho_2(a,b) \le \rho_1(a,b)$ it is useful to define the following combined distance $\rho_c(a,b) = (\rho_1(a,b) + \rho_{max}(a,b)) / 2$.

Note that 0-1 distance

$$\rho_{01}(a,b) = \begin{cases} 0 & if \quad a \equiv b \\ 1 & otherwise \end{cases}$$

defines also a metric in the plane.

5	5	5	5	5	5	5	5	5	5	5
5	4	4	4	4	4	4	4	4	4	5
5	4	3	3	3	3	3	3	3	4	5
5	4	3	2	2	2	2	2	3	4	5
5	4	3	2	1	1	1	2	3	4	5
5	4	3	2	1	0	1	2	3	4	5
5	4	3	2	1	1	1	2	3	4	5
5	4	3	2	2	2	2	2	3	4	5
5	4	3	3	3	3	3	3	3	4	5
5	4	4	4	4	4	4	4	4	4	5
5	5	5	5	5	5	5	5	5	5	5

Table 1: $\rho_{max}^{(5)}(a,b)$

					5					
			5.0	4.5	4	4.5	5.0			
		4.5	4.0	3.5	3	3.5	4.0	4.5		
	5.0	4.0	3.0	2.5	2	2.5	3.0	4.0	5.0	
	4.5	3.5	2.5	1.5	1	1.5	2.5	3.5	4.5	
5	4	3	2	1	0	1	2	3	4	5
	4.5	3.5	2.5	1.5	1	1.5	2.5	3.5	4.5	
	5.0	4.0	3.0	2.5	2	2.5	3.0	4.0	5.0	
		4.5	4.0	3.5	3	3.5	4.0	4.5		
			5.0	4.5	4	4.5	5.0			
					5					

Table 3: $\rho_c^{(5)}(a,b)$

					5					
				5	4	5				
			5	4	3	4	5			
		5	4	3	2	3	4	5		
	5	4	3	2	1	2	3	4	5	
5	4	3	2	1	0	1	2	3	4	5
	5	4	3	2	1	2	3	4	5	
		5	4	3	2	3	4	5		
			5	4	3	4	5			
				5	4	5				
					5					

Table 2: $\rho_1^{(5)}(a,b)$

					5					
		5.00	4.47	4.12	4	4.12	4.47	5.00		
	5.00	4.24	3.61	3.16	3	3.16	3.61	4.24	5.00	
	4.47	3.61	2.83	2.24	2	2.24	2.83	3.61	4.47	
	4.12	3.16	2.24	1.41	1	1.41	2.24	3.16	4.12	
5	4	3	2	1	0	1	2	3	4	5
	4.12	3.16	2.24	1.41	1	1.41	2.24	3.16	4.12	
	4.47	3.61	2.83	2.24	2	2.24	2.83	3.61	4.47	
	5.00	4.24	3.61	3.16	3	3.16	3.61	4.24	5.00	
		5.00	4.47	4.12	4	4.12	4.47	5.00		
					5					

Table 4: $\rho_2^{(5)}(a,b)$

This distance is called bounded because $\rho_{01}(a,b) \le 1 < \infty$ for every $a,b \in R^2$. We define bounded variants of other distances, namely

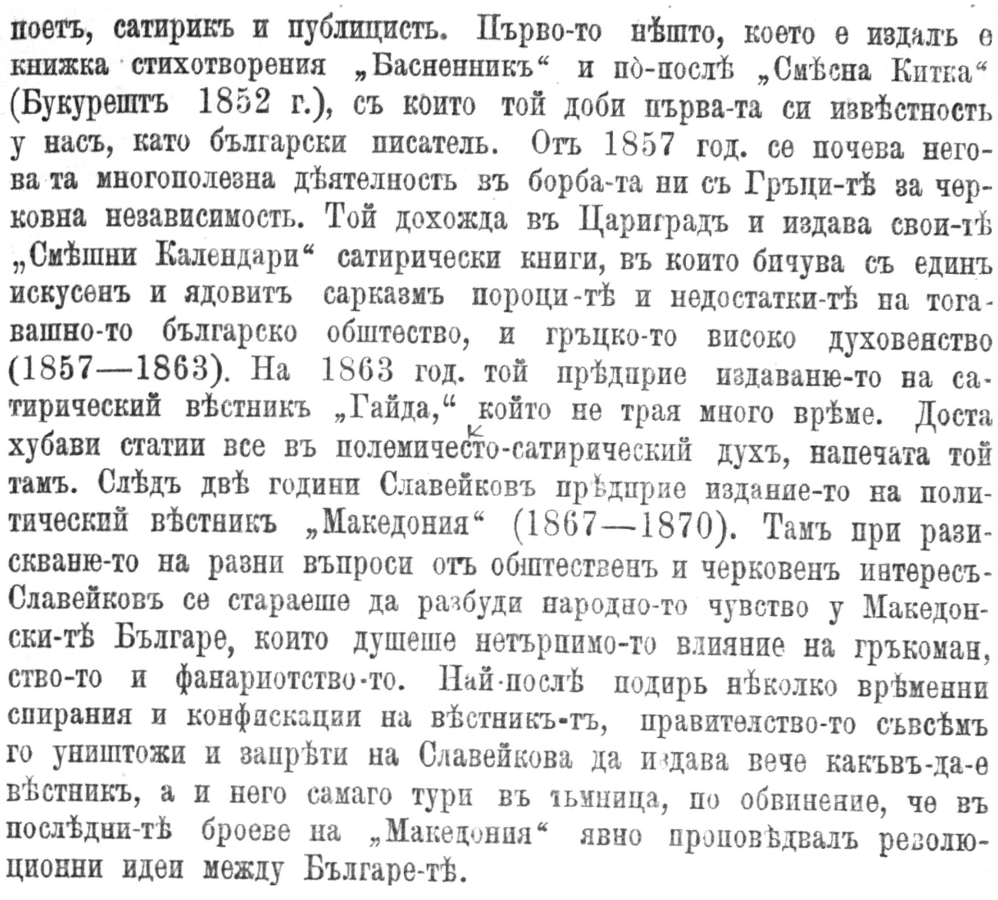
$$\rho^{(\tau)}(a,b) = \min\{\rho(a,b), \tau\},$$

where τ is a fixed positive number.

For integer net we calculate bounded distances from the origin to any point of the net for $\tau = 5$, see Tables 1-4.

4. EXPERIMENTS

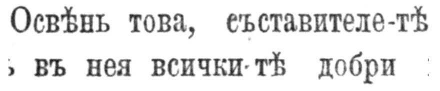
We carried out our experiments using an old book (1884) - Bulgarian Chrestomathy, created by famous Bulgarian writers Ivan Vasov and Konstantin Velichkov. We used 200 pages from 953 book's pages scanned at a resolution of 200 DPI as shown in Fig. 4.



поетъ, сатирикъ и публицистъ. Първо-то нѣшто, което е издалъ е книжка стихотворения „Басненникъ“ и пò-послѣ „Смѣсна Китка“ (Букурештъ 1852 г.), съ които той доби първа-та си извѣстность у насъ, като български писатель. Отъ 1857 год. се почева негова та многополезна дѣятелность въ борба-та ни съ Гърци-тѣ за черковна независимость. Той дохожда въ Цариградъ и издава свои-тѣ „Смѣшни Календари“ сатирически книги, въ които бичува съ единъ искусенъ и ядовитъ сарказмъ пороци-тѣ и недостатки-тѣ на тогавашно-то българско обштество, и гръцко-то високо духовенство (1857—1863). На 1863 год. той прѣдприе издаваню-то на сатирический вѣстникъ „Гайда,“ който не трая много врѣме. Доста хубави статии все въ полемичесто-сатирический духъ, напечата той тамъ. Слѣдъ двѣ години Славейковъ прѣдприе издание-то на политический вѣстникъ „Македония“ (1867—1870). Тамъ при разискваню-то на разни въпроси отъ обштественъ и черковенъ интересъ- Славейковъ се стараеше да разбуди народно-то чувство у Македонски-тѣ Българе, които душеше нетърпимо-то влияние на гръкоман, ство-то и фанариотство-то. Най-послѣ подирь нѣколко врѣменни спирания и конфискации на вѣстникъ-тъ, правителство-то съвсѣмъ го унищтожи и запрѣти на Славейкова да издава вече какъвъ-да-е вѣстникъ, а и него самаго тури въ тъмница, по обвинение, че въ послѣдни-тѣ броеве на „Македония“ явно проповѣдвалъ резолюционни идеи между Българе-тѣ.

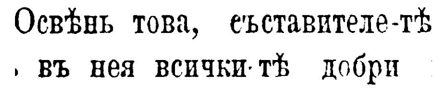
Figure 4: A half page of the book, grayscale

The goal of our experiments is to compare practically the efficiency of searching, counting the number of correctly retrieved words in a sequence of words, sorted by values of similarity measure. For all experiments the same segmentation parameters are used. We choose a pattern word and then measure similarities between it and the words with approximately same length.



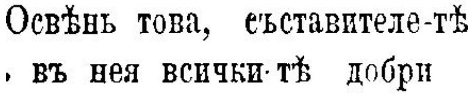
Освѣнь това, съставителе-тѣ
въ нея всички-тѣ добри

Figure 6: Original resolution, grayscale



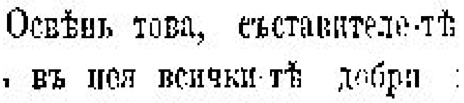
Освѣнь това, съставителе-тѣ
въ нея всички-тѣ добри

Figure 7: Original resolution, b/w



Освѣнь това, съставителе-тѣ
, въ нея всички·тѣ добри

Figure 8: Half resolution



Освѣнь това, съставителе-тѣ
, въ нея всички·тѣ добри :

Figure 9: Quarter resolution

For our experiments we choose a pattern word всички. There are two relative words (derivatives) of this word всичка and всичко. We count as correct all three of them.

In Tables 5 and 6 we count the number of correctly retrieved words among first 100, 200, 300, 400, 500 words with approximately same length.

$n =$	100	200	300	400	500
$\rho_{max}^{(5)}$	96	165	188	198	201
$\rho_1^{(5)}$	97	165	189	199	202
$\rho_2^{(5)}$	98	165	189	200	202
ρ_{01}	97	165	189	199	201

Table 5: Point distances,1170 x 1836 Pixels

$n =$	100	200	300	400	500
2340 x 3672 Pixels	100	177	205	213	220
1170 x 1836 Pixels	97	165	189	199	202
585 x 918 Pixels	93	139	157	168	174

Table 6: Resolution, $\rho_1^{(5)}$

Our results for using a number of point distances (see Table 5) manifest that all cases are practically identical. Table 6 shows that higher resolution does not involve essentially better results. In our example the middle table line ensure good retrieval combined with small execution time.

REFERENCES

[1] A. Andreev, N. Kirov, *Hausdorff Distance and Word Matching*, Proceedings of the International Workshop “Computer Science and Education”, June 3-5, 2005, Borovetz-Sofia, Bulgaria, 19-28.

[2] A. Andreev, N. Kirov, *Word image matching in Bulgarian historical documents*, Review of the National Center for Digitalization, 8, (2006), 29-35.

[3] A. Andreev, N. Kirov, *Text Search in Document Images Based on Hausdorff Distance Measures*, Proc. CompSysTech'08, 2008 (accepted).

[4] M.-P. Dubuisson, A. Jain, *A Modified Hausdorff Distance for Object Matching*, In: Proc. 12th Int. Conf. Pattern Recognition, Jerusalem, Israel, 1994, pp. 566-568.

[5] T. Konidaris, B. Gatos, K. Ntzios, I. Pratikakis, S. Theodoridis, S. J. Perantonis, *Keyword-guided word spotting in historical printed documents using synthetic data and user feedback*, International Journal of Document Analysis and Recognition, 9, (2007) 167–177.

[6] Dong-Gyu Sim, Oh-Kyu Kwon, and Rae-Hong Park, *Object Matching Algorithms Using Robust Hausdorff Distance Measures*, IEEE Trans. on Image Processing, 8, (1999), No.3, 425-429.

[7] Hwa-Jeong Son, Soo-Hyung Kim, Ji-Soo Kim, *Text image matching without language model using a Hausdorff distance*, to appear in: Information Processing and Management, (2008).