Serdica J. Computing **3** (2009), 23–46

Serdica
Journal of Computing

Bulgarian Academy of Sciences
Institute of Mathematics and Informatics

HAUSDORFF DISTANCES FOR SEARCHING IN BINARY TEXT IMAGES*

Andrey Andreev, Nikolay Kirov

Abstract. Hausdorff distance (HD) seems the most efficient instrument for measuring how far two compact non-empty subsets of a metric space are from each other. This paper considers the possibilities provided by HD and some of its modifications used recently by many authors for resemblance between binary text images. Summarizing part of the existing word image matching methods, relied on HD, we investigate a new similar parameterized method which contains almost all of them as particular cases . Numerical experiments for searching words in binary text images are carried out with 333 pages of old Bulgarian typewritten text, 200 printed pages of Bulgarian Chrestomathy from year 1884, and 200 handwritten pages of Slavonic manuscript from year 1574. They outline how the parameters must be set in order to use the advantages of the proposed method for the purposes of word matching in scanned document images.

1. Introduction. Optical character recognition (OCR) is a widely used approach for conducting text retrieval from scanned document images. It converts

ACM Computing Classification System (1998): I.5.4.

Key words: Hausdorff distance, binary text image, word matching.

*This work has been partially supported by Grant No. DO 02-275, 16.12.2008, Bulgarian NSF, Ministry of Education and Science.

text images into a text file, recognizing every letter or punctuation mark and mapping it to a number, which is called a code. The most often used codes are ASCII (one byte-code) and UTF-8 (two bytes code). This technique is well developed and has high accuracy leading to the relatively easy task for searching words in a text file.

Sometimes OCR is an impossible or a very difficult process requiring dictionaries in the corresponding language. Often human efforts are needed to correct OCR errors which is a quite tedious job. Here are some obstacles to successful OCR:

- the quality of page images: bad original source or bad scan process;
- language dependency: alphabet; old letters without the coding tables; old grammar, obsolete words, phrases and idioms; dictionaries; multi-lingual documents.

One of the main reasons for converting binary text images to text file is search. Searching in a text file is a well-known task – finding a sub-string in a string – and there are efficient algorithms for solving it. The solution is almost exact – the pattern string coincides with the result, or can be approximated when the goal is to avoid some grammar changes of the searched word. Of course the process is language dependent.

We suggest a different approach: instead of applying two steps – OCR and searching in text documents, it is possible for words to be searched directly in scanned text documents (text images) (see [1]–[4]). Organizing retrieval of words, similar to a given pattern word, by searching in the set of binary text images is the idea suggested also in [14], [8], [11] and [17].

The goals of this paper are:

- to propose a new method for estimating the similarity between binary images in order to generalize and to unify the existing image matching methods based on Hausdorff distance;
- to check numerically the efficiency of the generalized HD method when it is applied for word matching in typewritten, printed and handwritten historical documents of bad quality, and, using the numerical results, to determine the values of the parameters on which it depends;
- to present and test numerically a convenient computer system which is practically useful in the word retrieval process.

2. Hausdorff distances for measuring set similarities. The Hausdorff distance (HD) between two closed and bounded subsets A and B of a given metric space M is defined by

$$H(A,B) = \max\{h(A,B), h(B,A)\}, \tag{1}$$

where $h(A,B)$ is so-called directed distance from A to B. For classical Hausdorff distance

$$h(A,B) = \max_{a\in A} d(a,B), \text{ where } \quad d(a,B) = \min_{b\in B} \rho(a,b). \tag{2}$$

$d(a,B)$ is the distance from a point a to the set B, and $\rho(a,b)$ is a point distance in the metric space M. The HD defined by (1) and (2) satisfies all metric requirements: $H(A,B) \geq 0$, $H(A,B) = 0 \;\Leftrightarrow\; A \equiv B$, $H(A,B) = H(B,A)$ and $H(A,B) \leq H(A,C) + H(B,C)$ for any subsets $A, B, C \subseteq M$.

HD looks very attractive for measuring the similarity between plane sets. Unfortunately, despite its metric properties mentioned above, the HD (1) does not meet robust requirements. Simple examples like the one given in [15] show that $H(A,B)$ could be a big number despite "visual" similarity between the sets A and B. Many attempts were made to avoid this "weakness" of HD modifying it in a way to overcome the representation of HD by just two points which could be parasitic (not part of a real image). The main idea is that at the expense of the loss of some metric requirements, for example triangle inequality or symmetry, more points have to be included, decreasing in such way the influence of the eventual presence of noise upon the final evaluation of $H(A,B)$.

For raster sensing devices it is enough for finite point sets to be considered. For any such set A on the plane let N_A denote its number of points.

D. P. Huttenlocher *et al.* [9] proposed the Partial Hausdorff Distance (PHD) for comparing images containing much degradation or occlusions. Let $K^{th}_{a\in A}$ denote the K-th ranked value in the set of distances $\{d(a,B) : a \in A\} = \{d(a_i,B), i = 1,\ldots,N_A\}$, i.e. for each point of A, the distance to the closest point of B is computed, and then, the points of A are ranked by their respective distance values,

$$d(a_1,B) \geq d(a_2,B) \geq \cdots \geq d(a_K,B) \geq \cdots \geq d(a_{N_A},B). \tag{3}$$

Let us note that our definition of $K^{th}_{a\in A}$ differs from the original one in [9], where the rating order in (3) is in the opposite direction. The directed distance for PHD is

$$h_K(A,B) = K^{th}_{a\in A} d(a,B) = d(a_K,B). \tag{4}$$

A parameter $\alpha = K/N_A$, $(0 < \alpha \leq 1)$ can be defined, relative to the set A. D.-G. Sim *et al.* [16] claim that $\alpha \approx 0.4$ provides good matching image results.

The idea of R. Azencott *et al.* [5], and J. Paumard [15] is that we do not take into account the L closest neighbors of $a \in A$ in B. So we define the distance from a point $a \in A$ to the set B as follows

$$d_L(a, B) = L^{th}_{b \in B} \rho(a, b),$$

where $L^{th}_{b \in B} \rho(a, b) = \rho(a, b_L)$ denotes the L-th ranked value in the set of distances $\{\rho(a, b) : b \in B\} = \{\rho(a, b_i), i = 1, \ldots, N_B\}$, i.e.

$$\rho(a, b_1) \leq \cdots \leq \rho(a, b_L) \leq \cdots \leq \rho(a, b_{N_B}).$$

Note that if $L > 1$, $A \equiv B$ and $a \in A$ then $d_L(a, A) > 0$. Now the directed Censored Hausdorff Distance (CHD) is defined by

$$h_{\alpha,\beta}(A, B) = h_{K,L}(A, B) = K^{th}_{a \in A} d_L(a, B) = K^{th}_{a \in A} L^{th}_{b \in B} \rho(a, b). \tag{5}$$

The parameters $\alpha = K/N_A$ and $\beta = L/N_B$ are relative values with respect to the number of points in the sets A and B. For comparing images obtained by adding randomly black and white dots to one of them the recommended values in [15] for the parameters are $\alpha = 0.1$ and $\beta = 0.01$. Evidently CHD does not meet identity and triangle inequality metric properties.

M.-P. Dubuisson and A. Jain [7] examined 24 distance measures of Hausdorff type for determining to what extend two finite sets A and B on the plane differ. Based on the numerical behavior of these distances on synthetic images containing various levels of noise they introduced the so-called Modified Hausdorff Distance (MHD), whose directed distance is

$$h_{\mathrm{MHD}}(A, B) = \frac{1}{N_A} \sum_{a \in A} d(a, B) = \frac{1}{N_A} \sum_{a \in A} \min_{b \in B} \rho(a, b). \tag{6}$$

They claim than it suit best the matching problem for noisy images supposing that ρ is the Euclidean metric (12). Applying MHD we use the point distance (14) for our experiments in [1]–[4] measuring the word similarities in binary text images and conclude that this is one of the best measures for word matching.

A similar approach called Weighted Hausdorff Distance is used in [14] for finding a word image matching method in English and Chinese document images. The authors suppose that the contribution of different parts of the word image to HD is not the same.

Word matching numerical experiments with typewritten Bulgarian historical documents show that somewhat better results were obtained omitting the coefficient $1/N_A$ in front of the sum (6). We called (see [4]) this tiny modification Sum Hausdorff Distance (SHD) with directed distance

$$h_{\mathrm{SHD}}(A,B) = \sum_{a\in A} d(a,B) = \sum_{a\in A} \min_{b\in B} \rho(a,b). \tag{7}$$

In 1999 D.-G. Sim *et al.* [16] described two variants of MHD for elimination of outliers – usually the points of outer noise. Based on robust statistics, M-estimation and least trimmed square, they introduced M-HD and LTS-HD. The directed M-HD is defined by

$$h_{\mathrm{M}}(A,B) = \frac{1}{N_A} \sum_{a\in A} f(d(a,B)), \tag{8}$$

where the function f is symmetric and has an unique minimum value at zero. They introduce one simple function with these properties:

$$f(x) = \begin{cases} |x| & \text{if } |x| \le \tau, \\ \tau & \text{if } |x| > \tau. \end{cases} \tag{9}$$

The recommended interval of τ is $[3,5]$ for their purposes.

The second measure proposed in [16] is called Least Trimmed Square HD (LTS-HD) with directed distance

$$h_{\mathrm{LTS}}(A,B) = \frac{1}{N_A - K + 1} \sum_{i=K}^{N_A} d(a_i, B), \tag{10}$$

where $1 \le K \le N_A$ and $a_1, a_2, \ldots, a_{N_A}$ are all points of A for which (3) is valid. The parametrization of the method can be done by a parameter $\alpha = K/N_A$. The suggested value α for comparing noisy binary images contaminated by Gaussian noise is 0.2.

In 2008 E. Baudrier, *et al.* [6] try to avoid the noise in the images by means of Windowed Hausdorff Distance (WHD). They also defined a Local Distance Map:

$$LDM(x) = \begin{cases} d(x,A) & \text{if } x \notin A, x \in B; \\ d(x,B) & \text{if } x \in A, x \notin B; \\ 0 & \text{otherwise,} \end{cases} \tag{11}$$

where d is defined by (2). The function $LDM(x)$ provides excellent visual representation of the local differences between the sets A and B. For quantitative comparing of word images we need numbers, and for converting LDM to a number, we could apply an appropriate norm to the function $LDM(x)$. Note that it is fulfilled that

$$H(A,B) = \sup_{x\in A\cup B} LDM(x) \quad \text{and}$$

$$H_{SHD}(A,B) = \max\left\{\sum_{a\in A} LDM(a), \sum_{b\in B} LDM(b)\right\} \leq \sum_{x\in A\cup B} LDM(x).$$

3. Point distances. The Minkowski distance of order $p \geq 1$ between two points $a = (a_x, a_y)$ and $b = (b_x, b_y)$ in the plane is defined as:

$$\rho_p(a,b) = \left(|a_x - b_x|^p + |a_y - b_y|^p\right)^{1/p}.$$

The following distances will be used:

$$\text{Euclidean distance}: \quad \rho_2(a,b) = \sqrt{(a_x-b_x)^2 + (a_y-b_y)^2}. \tag{12}$$

$$\text{Manhattan distance}: \rho_1(a,b) = |a_x - b_x| + |a_y - b_y|. \tag{13}$$

$$\text{Chebyshev distance}: \quad \rho_\infty(a,b) = \rho_{\max}(a,b) = \max\{|a_x - b_x|, |a_y - b_y|\}. \tag{14}$$

In practice of image comparison, we have an upper limit for distances between the points of the images. That is why we define bounded variants of point distances:

$$\rho^{(\tau)}(a,b) = \min\{\rho(a,b), \tau\}, \tag{15}$$

where τ is a positive number and $\rho(a,b)$ can be any distance. The point distance (15) is a metric for any $\tau > 0$. For raster images (pixels are defined on a quadratic net with a side lenght equal to 1) in the case of $\tau = 1$ we have:

$$\rho_1^{(1)}(a,b) = \rho_2^{(1)}(a,b) = \rho_{\max}^{(1)}(a,b) = \begin{cases} 0 & \text{if } a \equiv b \\ 1 & \text{otherwise} \end{cases} \tag{16}$$

In the above notations M-HD (8) with the function (9) coincides with MHD (6) with the point distance $\rho^{(\tau)}$.

When bounded distances between image pixels are calculated, $(\tau+1) \times (\tau+1)$ matrices can be created initially (see [12]) in order to unify the usage of all three norms (13)–(14) and to avoid multiplication and square root for $\rho_2^{(\tau)}$.

4. A new approach to HD similarity measures. An idea for generalizing of HD distances was given in [3]. Let A and B be finite sets on the plane and let us suppose a linear order of the points of A:

$$A = \{a_1, a_2, \ldots, a_{N_A}\}.$$

For every $a_k \in A$ $(k = 1, \ldots, N_A)$ we calculate the distances from a_k to all points in B, with respect to a given metric $\rho = \rho^{(\tau)}$ defined by (15), as follows:

$$d_{k1} = \min\nolimits_{b\in B} \rho(a_k, b) = \rho(a_k, b_{k1}), \quad d_{k2} = \min\nolimits_{b\in B\setminus\{b_{k1}\}} \rho(a_k, b) = \rho(a_k, b_{k2}), \ldots,$$
$$d_{kl} = \min\{\rho(a_k, b) : b \in B\setminus\{b_{k1}, b_{k2}, \ldots, b_{kl-1}\}\} = \rho(a_k, b_{kl}), \ldots,$$
$$d_{kN_B} = \min\{\rho(a_k, b) : b \in B\setminus\{b_{k1}, b_{k2}, \ldots, b_{kl-1}, \ldots, b_{kN_B-1}\}\} = \rho(a_k, b_{kN_B}).$$

In such a way we obtain a nondecreasing sequence of nonnegative numbers

$$d_{k1} \le d_{k2} \le \cdots \le d_{kl} \le \cdots \le d_{kN_B}. \tag{17}$$

Let the matrix D be defined by

$$D = \begin{pmatrix} d_{11} & d_{12} & \ldots & d_{1l} & \ldots & d_{1N_B} \\ d_{21} & d_{22} & \ldots & d_{2l} & \ldots & d_{2N_B} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ d_{k1} & d_{k2} & \ldots & d_{kl} & \ldots & d_{kN_B} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ d_{N_A1} & d_{N_A2} & \ldots & d_{N_Al} & \ldots & d_{N_AN_B} \end{pmatrix}$$

For every $1 \le l \le N_B$, we also define a matrix D_l interchanging the rows of the matrix D

$$D_l = \begin{pmatrix} d^l_{11} & d^l_{12} & \ldots & d^l_{1l} & \ldots & d^l_{1N_B} \\ d^l_{21} & d^l_{22} & \ldots & d^l_{2l} & \ldots & d^l_{2N_B} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ d^l_{k1} & d^l_{k2} & \ldots & d^l_{kl} & \ldots & d^l_{kN_B} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ d^l_{N_A1} & d^l_{N_A2} & \ldots & d^l_{N_Al} & \ldots & d^l_{N_AN_B} \end{pmatrix} \tag{18}$$

so that the elements of l-th column are sorted, i.e. satisfy the following inequalities:

$$d^l_{1l} \ge d^l_{2l} \ge \cdots \ge d^l_{kl} \ge \cdots \ge d^l_{N_Al}. \tag{19}$$

Let $0 < k \le N_A$ and $0 < l \le N_B$ be integer numbers. We define two Generalized Hausdorff Distances (GHD):

$$h^{(p)}_{k,l}(A, B) = d^l_{kl}, \quad H^{(p)}_{k,l}(A, B) = \max\{h^{(p)}_{k,l}(A, B), h^{(p)}_{k,l}(B, A)\} \tag{20}$$

and

$$
\begin{aligned}
h_{k,l}^{(s)}(A,B) &= \frac{1}{N_A-k+1}\sum_{i=k}^{N_A} d_{il}^{l}, \\
H_{k,l}^{(s)}(A,B) &= \max\{h_{k,l}^{(s)}(A,B), h_{k,l}^{(s)}(B,A)\}.
\end{aligned}
\tag{21}
$$

We call $H_{k,l}^{(p)}$ p-distance and $H_{k,l}^{(s)}$ s-distance.

The definitions (20) and (21) are generalizations of all Hausdorff based distances mentioned in Section 2, which can be represented by

1. directed HD (2): $h(A,B) = h_{1,1}^{(p)}(A,B) = d_{11}^{1}$;
2. directed PHD (4): $h_K(A,B) = h_{K,1}^{(p)}(A,B) = d_{K1}^{1}$;
3. directed CHD (5): $h_{K,L}(A,B) = h_{K,L}^{(p)}(A,B) = d_{KL}^{L}$;
4. directed MHD (6): $h_{\mathrm{MHD}}(A,B) = h_{1,1}^{(s)}(A,B) = \dfrac{1}{N_A}\displaystyle\sum_{i=1}^{N_A} d_{i1}^{1}$;
5. directed SHD (7): $h_{\mathrm{SHD}}(A,B) = N_A.h_{\mathrm{MHD}}(A,B) = N_A.h_{1,1}^{(s)}(A,B) = \displaystyle\sum_{i=1}^{N_A} d_{i1}^{1}$;
6. directed LTS-HD for a given $1 \le K \le N_A$ (10): $h_{\mathrm{LTS}}(A,B) = h_{K,1}^{(s)}(A,B)$.

Now we change the parametrization of (20) and (21) replacing k and l by parameters α and β relative to the sets A and B:

$$
\alpha = \frac{k-1}{N_A} \quad \text{and} \quad \beta = \frac{l-1}{N_B}. \tag{22}
$$

Since $1 \le k \le N_A$ and $1 \le l \le N_B$ we have $\alpha, \beta \in [0,1)$. So for defining a concrete generalized HD, we have to take a vector of parameters $(\alpha, \beta, \tau, \rho)$ and chose p- or s-distance.

Local Distance Map (11) can be generalized too. Let us choose β, τ, and ρ. If $x = a_k \in A$ in the order (17), then $GLDM(x) = d_{kl}$, where $l = \beta N_B$, while $LDM(x) = d_{k1}$. Symmetrically, if $x = b_k \in B$ a sequence like (17) can be create and we set $GLDM(x) = d_{kl}$, where d_{kl} is l-th distance from b_k to the set A and $l = \beta N_A$. If $x \notin A \cup B$, then $GLDM(x) = LDM(x) = 0$.

5. Measuring the effectiveness of searching. The effectiveness of searching methods is usually given by standard estimations *Recall* and *Precision* (see M. Junker *et al.* [10]). Let us look for a word W_0 (pattern word) in a collection of binary text images in which W_0 occurs N times. Comparing W_0 with other words in the text, a sequence of words is generated:

$$\{W_i\}_{i=0,1,\ldots} \tag{23}$$

ordered according to a similarity measure based on some HD.

For every positive integer n, let $m(n) \leq n$ be the number of words among the first n words of (23) that coincide with W_0 as words. Obviously $0 \leq m(n) \leq N$. Then *Recall* $r(n)$ and *Precision* $p(n)$ are defined by

$$r(n) = \frac{m(n)}{N} \quad \text{and} \quad p(n) = \frac{m(n)}{n}. \tag{24}$$

The function

$$P : [0,1] \to [0,1], P(r) = P(r(n)) = p(n) \tag{25}$$

represents the effectiveness of searching the word W_0. In the ideal case, when the first N words in the sequence (23) are correct, the function $P(r(n))$ is the constant 1, because $m(n) = n$ for all $n \leq N$. When we compare GHD with different parameters, bigger values of P means better choice of the parameters. Also the number $r_1 \in [0,1]$,

$$r_1 = \max\{r(n) : p(n) = 1\} \tag{26}$$

is an important characteristic of the applied GHD measure.

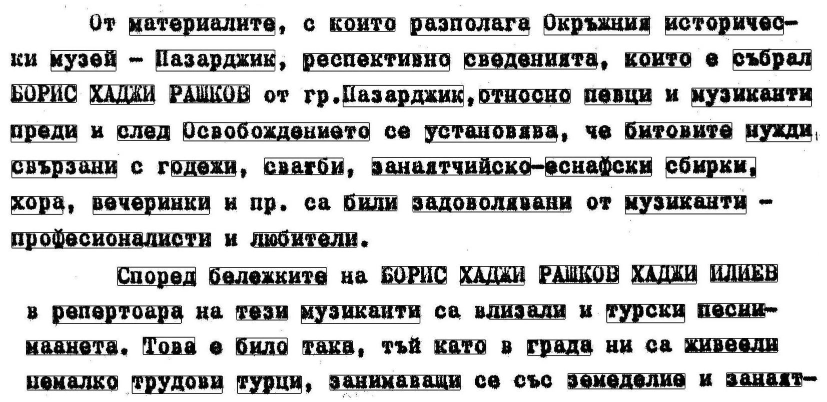
6. Experiments. The main objectives of computer experiments are:

- to prove that our approach gives good results in practice;
- to compare search results with respect to parameter values of GHD using different types of documents – printed, typewritten and handwritten.

We apply p-distance (20) or s-distance (21) under parametrization (22). In the next when we mention p-distance, this means that the sorting algorithm for producing the word sequence (23) uses the primary sort key (20) and secondary sort key (21) and vice versa. This approach avoids the discontinuity of p-distance (see [1], [4] and [2]) when the words in the sequence (23) are divided into a several classes, corresponding to the distance to the pattern. This effect is stronger for $\rho_{\max}$ and ρ_1 point distances but can be manifested for ρ_2 too.

For every instance we choose parameters $\alpha \in [0, 0.5]$, $\beta \in [0, 0.05]$, $\tau \in \{1, 2, \ldots, 19\}$, and $\rho \in \{\rho_1, \rho_2, \rho_{\max}\}$.

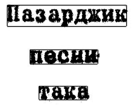
6.1. Typewritten text. Bulgarian typewritten text of 333 poor quality pages is the data in our experiments (see book [20]). This text has also been used in [1], [4] and [2]. The pages are in `tif` format, black and white (1 Bpp), approximately 2300×3300 pixels (Fig. 1). We use three words of different length for searching the document. All pattern words (see Fig. 2) are chosen from the first page, no special requirements are imposed on them. From the users point of view, this is the most natural choice.



От материалите, с които разполага Окръжния историчес-
ки музей - Пазарджик, респективно сведенията, които е събрал
БОРИС ХАДЖИ РАШКОВ от гр.Пазарджик,относно певци и музиканти
преди и след Освобождението се установява, че битовите нужди,
свързани с годежи, сватби, занаятчийско-еснафски сбирки,
хора, вечеринки и пр. са били задоволявани от музиканти -
професионалисти и любители.

Според бележките на БОРИС ХАДЖИ РАШКОВ ХАДЖИ ИЛИЕВ
в репертоара на тези музиканти са влизали и турски песни-
маанета. Това е било така, тъй като в града ни са живеели
немалко трудови турци, занимаващи се със земеделие и занаят-

Fig. 1. As segmented fragment from the first page



Пазарджик

песни

така

Fig. 2

The relatively long (9 letters) word **Пазарджик** (Pazardzhik the name of a Bulgarian town) is a pattern word for this experiment. It occurs 231 times in the text but the number of correctly segmented words **Пазарджик** is 200, so we set $N = 200$ (see Section 5).

Fig. 3 presents 5 graphics of functions P defined by (25) and obtained by GHD as s-distance with parameters $\alpha = 0, 0.01, 0.05$ and 0.08, $\beta = 0.005$ and $\rho^{(\tau)} = \rho^{(15)}_{\max}$. If a parameter is fixed for all graphics in a figure, then it is set in the caption of the figure, otherwise it is set in the chart legend. Fig. ?? shows that $\alpha = 0.01$ gives the maximum of $r_1 = 0.805$, see (26). This result is obtained when $n = 161$ and of course $m(161) = 161$ which means that the first 161 words in the sequence (23) are correct. But this α value does not give the best result in the entire interval $[0, 1]$ because $\alpha = 0.03$ reaches maximum of $r(326) = 0.97$ with the best precision $p(326) = 0.595$. Note than the function P is not defined in the interval $(0.97, 1]$ because $m(500) = m(326) = 194$ while $N = 200$.

The results presented in Fig. 4 show that $\beta = 0.001$ is the best choise of an s-distance among the values 0, 0.001, 0.005 and 0.01 when $\alpha = 0$, $\tau = 15$ and $\rho = \rho_{\max}$.

The influence of τ is not essential (Fig. 5), nevertheless $\tau = 1$ does not

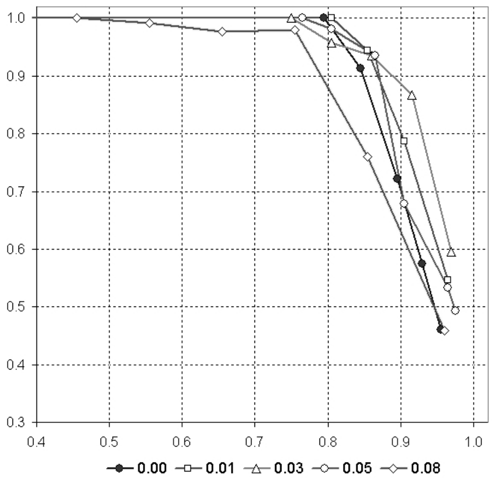


Fig. 3. **Пазарджик**: $(\alpha, 0.005, 15, \rho_{\max})$,s

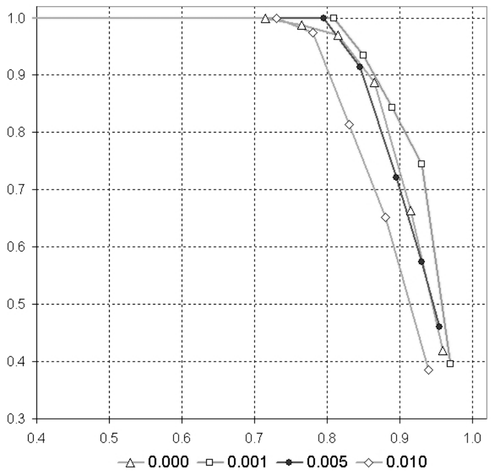


Fig. 4. **Пазарджик**: $(0, \beta, 15, \rho_{\max})$,s

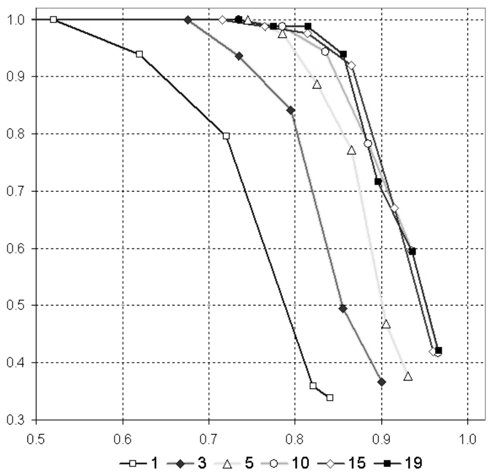


Fig. 5. **Пазарджик**: $(0, 0, \tau, \rho_{\max})$,s

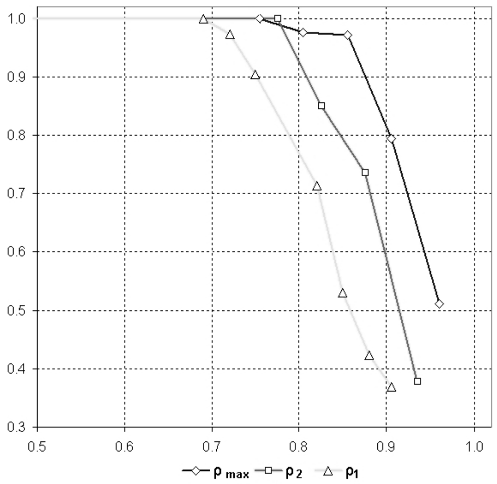


Fig. 6. **Пазарджик**: $(0.02, 0.005, 15, \rho)$,s

ensure a satisfactory result, while the graphics for $\tau = 15$ and $\tau = 19$ are above other graphics and hence give a better result. The $\rho_{\max}$ point distance is better than ρ_1 and ρ_2 in the case while $\alpha = 0.02$, $\beta = 0.005$ and $\tau = 15$ (Fig. 6).

Fig. 7 shows that if $(\alpha, \beta) = (0, 0)$ then the s-distance is better than p-distance while for $(\alpha, \beta) = (0.03, 0.005)$ the graphic of the p-distance is placed above the graphic of the s-distance when $r(n) \in [0.825, 0.945]$. The best results

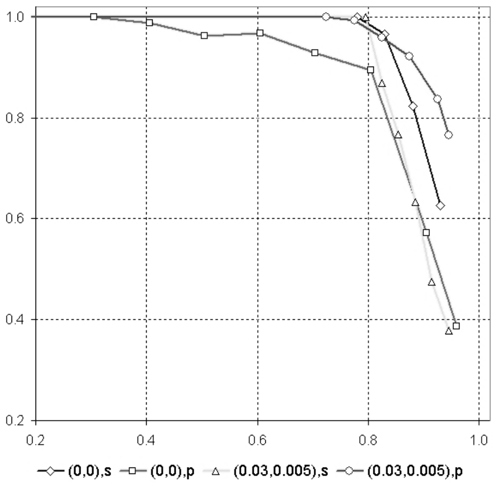


Fig. 7. **Пазарджик**: $(\alpha, \beta, 19, \rho_2)$

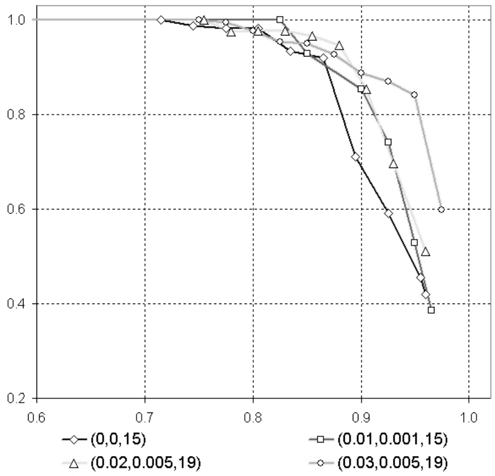


Fig. 8 **Пазарджик**: $(\alpha, \beta, \tau, \rho_{\max})$,s

for word **Пазарджик**, obtained in our experiments for s-distance and $\rho = \rho_{\max}$, are given in Fig. 8. We see that there is no best set of parameters – the maximum $r(n) = r_1 = 0.825$ for $p(n) = 1$ is reached for $(0.01, 0.001, 15)$ while for $r(n) \in [0.9, 0.975]$ the best parameter set is $(0.03, 0.005, 19)$.

On Fig. 9 the pattern word **Пазарджик** is placed in the middle of the picture. The corresponding left and right words are compared with the pattern word and the corresponding *LDM* function gives a visual notion of the differences in the words. Here $\rho^{(1)}$ norm is applied and therefore the *LDM* image (the black dots) is $(A \backslash B) \cup (B \backslash A)$.

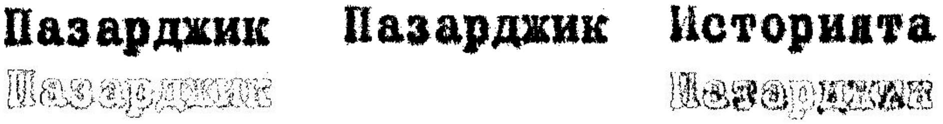


Fig. 9. **Пазарджик**: LDM

Fig. 10 represents the results for the word **песни** (songs) – 70 times in the text. In this case there is a best set of parameters, namely $(0.01, 0.001, 19, \rho_2)$,p. Also p-measures are better than the corresponding s-measures for every $r(n)$.

From a practical point of view, the list of words (23) obtained by the software, contains additional information – words similar (as images) to the pattern, which are closely related (of meaning) to this word: **песни** (songs), **песен** (song),

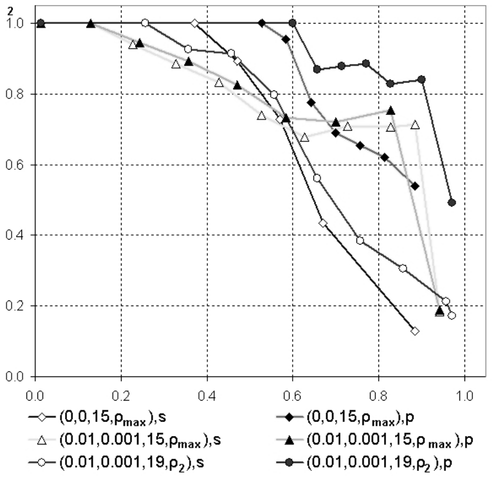


Fig. 10. **песни**: the best

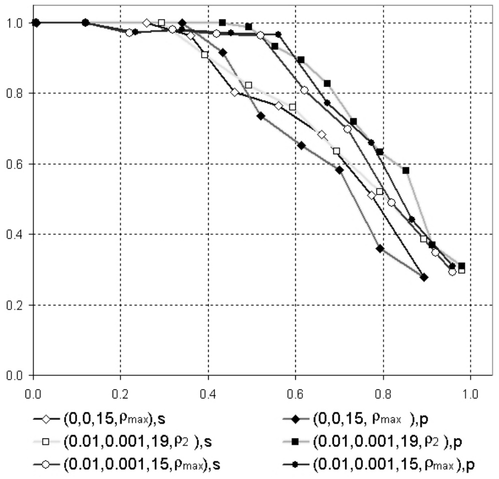


Fig. 11. **песни** & relatives

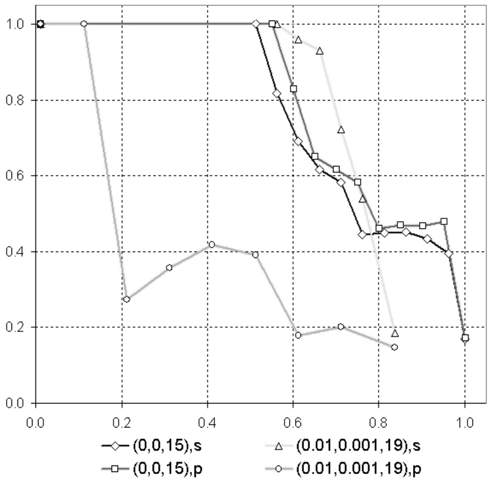


Fig. 12. **така**: $(\alpha, \beta, \tau, \rho_{\max})$

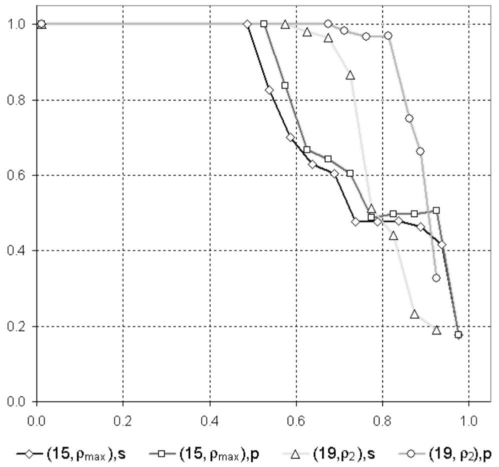


Fig. 13. **така**: $(0.01, 0.001, \tau, \rho)$

пеене (singing), **пеели** (sang), **пеела** (sang), **пеело** (sang), **пееше** (sang), **пееха** (sang), **певни** (in a sing-song manner), **певец** (singer), **певци** (singers). We called these words relatives. This is a common situation for the Bulgarian language because many grammatical changes of a given word produce similar looking words changing only one or two letters. We can consider this fact as an additional advantage to the user in searching process. The recall-precision graphics for relative words can be seen on Fig. 11. To create these graphics we should count the number N of relatives but we prefer to estimate it, setting $N = 150$ because $p(500) = 147$ in the best case. All three sets of parameters provide good practical results.

Figures 12 and 13 represent the results for the pattern word **така** (thus, like this/that, so, sure). It occurs 80 times in the text. There are several four-letter words, very similar (as images) to the pattern word: **това**, **този**, **тази** (this), **тези** (these), etc. which are contained in the text many times. These words have different meaning and the user cannot utilizes them (as in the case of **песни**).

The best result is obtained by p-distance with parameters $(0.01, 0.001, 19, \rho_2)$. 67,5% of the words are retrieved with precision 100% ($r_1 = 0.675$) and 81% with precision 97%. Also $m(70) = 67$, i.e. only 3 incorrect words can be found in the first 70 members of sequence (23).

GLDM in Fig. 14 gives an idea of a case when $H^{(s)}_{1,1}(W_0, W_{i_1}) > H^{(s)}_{1,1}(W_0, W_{i_2})$ (and hence $i_1 > i_2$) but W_{i_1} is a correct word while W_{i_2} is incorrect one. Here W_{i_1} = "**така**", W_{i_2} = "**това**", $i_1 = 359$, $i_2 = 48$, $H^{(s)}_{1,1}(W_0, W_{i_1}) = 35.9$, $H^{(s)}_{1,1}(W_0, W_{i_2}) = 23.4$. The numbers are obtained using $\beta = 0$, $\tau = 15$ and $\rho = \rho_2$.

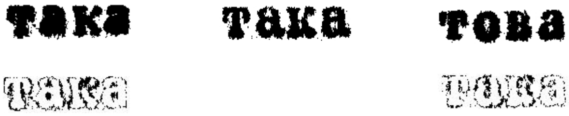


Fig. 14. **така**: LDM $(0, 15, \rho_2)$

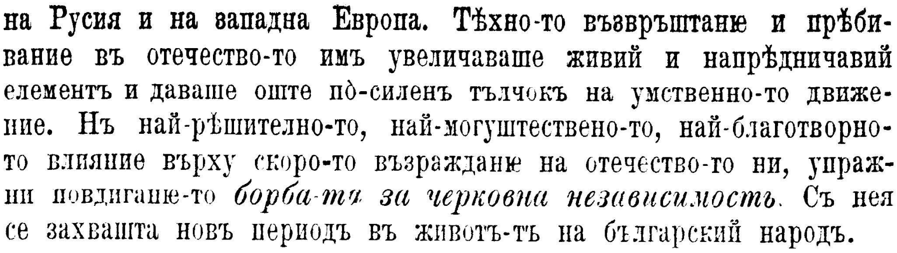
6.2. Printed text. The experiments carried out are using an old book (1884) – a Bulgarian Chrestomathy, composed by the famous Bulgarian writers Ivan Vazov and Konstantin Velichkov (see [19]).

Theoretically we can find all words in the printed text which coincide with a given pattern word (as by the operation "find" in a text file) under the assumption that scanned images are perfect. Often this is not the case. In this instance the quality of scanned images is quite bad because this was one of the first books, processed in the digitization center and operators' qualification was not on appropriate level. Many pages have slopes in rows, there are significant variations in gray levels, etc.

There is no text version till now of this book, which might be produced using appropriate OCR software. The first reason is the quality of images. The second reason is the absence of OCR software because the text contains obsolete

Bulgarian letters. Also spelling and grammar are quite different in modern Bulgarian language.

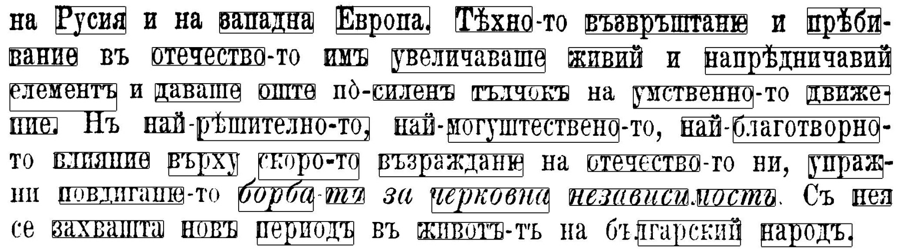
We used 200 images from about 1000 scanned pages, `tif` format, resolution about 2300×3800, 1 Bpp (see Fig. 15). The same pages from this book were used as data in articles [3] and [12].



на Русия и на западна Европа. Тѣхно-то възвръштанѥ и прѣбивание въ отечество-то имъ увеличаваше живий и напрѣдничавий елементъ и даваше оште пò-силенъ тълчокъ на умственно-то движение. Нъ най-рѣшително-то, най-могуштествено-то, най-благотворно-то влияние върху скоро-то възражданѥ на отечество-то ни, упражни повдиганѥ-то *борба-тѫ за черковна независимость*. Съ нея се захвашта новъ периодъ въ животъ-тъ на българский народъ.

Fig. 15. A fragment of Chrestomathy page

For our experiments we choose a pattern word **всички** (all, everyone). It is tedious to count all words **всички** in all 200 pages, but we can estimate this number quite precisely. The best result gives us 114 correct words in the first 500 of the sequence (23). The total number of checked words is 7505 and the distribution of correct words (see Table 1) is a reason for setting $N = 120$ and using this number in formulas (24).



на Русия и на западна Европа. Тѣхно-то възвръштанѥ и прѣбивание въ отечество-то имъ увеличаваше живий и напрѣдничавий елементъ и даваше оште пò-силенъ тълчокъ на умственно-то движение. Нъ най-рѣшително-то, най-могуштествено-то, най-благотворно-то влияние върху скоро-то възражданѥ на отечество-то ни, упражни повдиганѥ-то *борба-тѫ за черковна независимость*. Съ нея се захвашта новъ периодъ въ животъ-тъ на българский народъ.

Fig. 16. A fragment of segmented Chrestomathy page

Fig. 17 presents the results for s- and p-distances with a parameter $\rho \in \{\rho_{\max}, \rho_2\}$. For ρ_2, p-distance $r_1 > 0.5$, but for $\rho_{\max}$, s-distance the maximum retrieval of 0.95 is achieved. There are two relative words (derivatives) **всичка** (all, everyone) and **всичко** (all, everything) of the base word **всички**. We count as correct all three of them and show the results in Fig. 18 using $N = 230$. Now $\rho_{\max}$ with p-distance provides the best result $r_1 = 0.791$. The results for $r(n) > 0.85$ do not depend substantially on the plane metric.

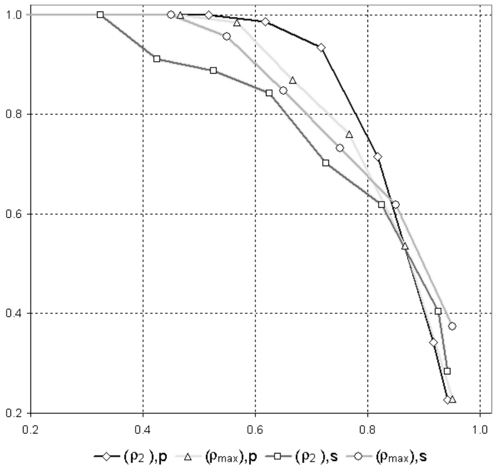


Fig. 17. **всички**: $(0.01, 0.001, 19, \rho)$

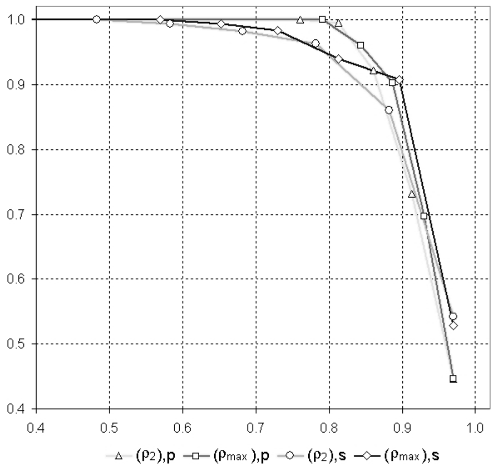


Fig. 18. **всички** & rel.: $(0.01, 0.001, 19, \rho)$

In Tables 1 and 2 we count the number of correctly retrieved words among first 100, 200, 300, 400, 500 words in the sequence (23), chosen from 7505 words with approximately same length. The parameters are: $(0.001, 0.001, 19, \rho_2)$, p-distance. The last row of both tables show the number of correct words in the interval $(n-100, n]$.

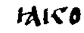
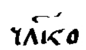
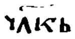
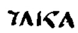
n	100	200	300	400	500
$m(n)$	89	106	109	111	112
	89	17	3	2	1

Table 1. **всички**

n	100	200	300	400	500
$m(n)$	100	192	210	218	223
	100	92	22	8	5

Table 2. **всички**, **всичка**, **всичко**

6.3. Handwritten text. The text under investigation is a Slavonic manuscript (Fig. 19), the Zbornik "Zlatoust" (1574) [18], 747 pages, `jpg`, resolution about 1250×1900 pixels, 24 Bpp (originally), converted to 1 Bpp `tif` images. We consider 200 pages for experiments. The segmentation is quite good because the writer have been clerkly hand and any simple algorithm could separate rows and words (see Fig. 20). The only problem in this step is a 1–2° slopes for almost all pages.

The pattern word  is located on page 8. In the actual Bulgarian alphabet it spells **яко**; in the text it has the obsolete meaning **като**, **както** (like, as). The same word is also written as . We count both words as correct retrievals. There are two more words  and  which are very similar (as

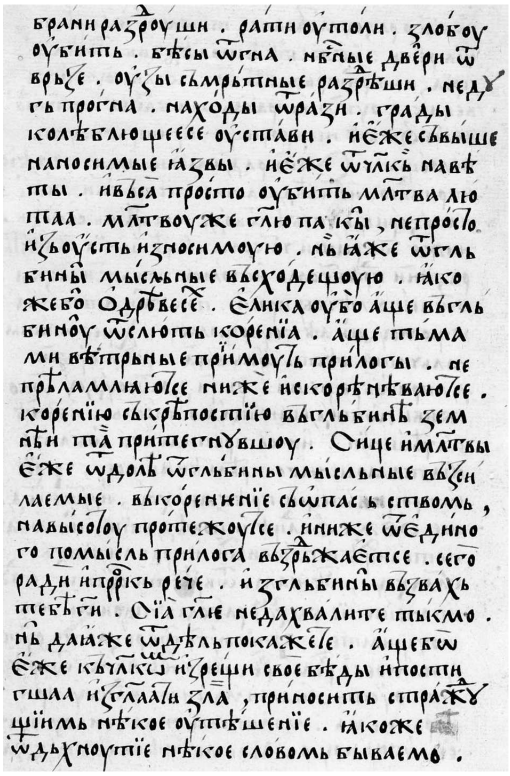


Fig. 19. Page 8

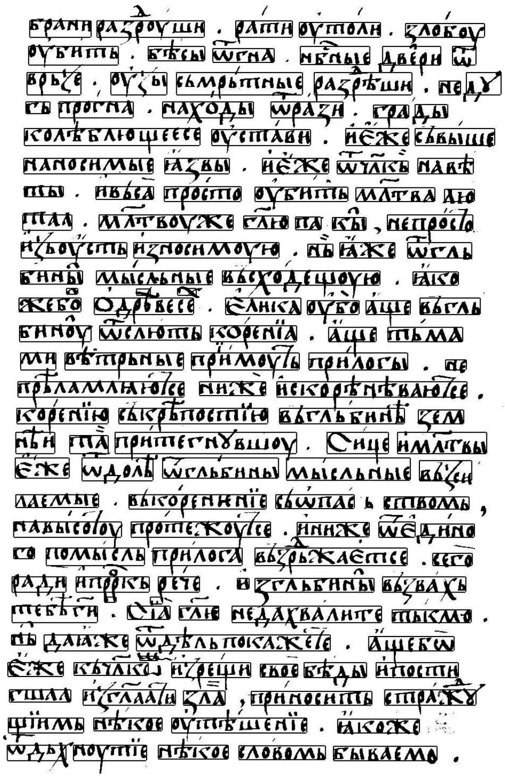


Fig. 20. Segmented page 8

images) but have different meanings and we do not count them as correct.

All experiments with reasonable parameters good for typewritten text (see Section 6.1), produced very similar results. We do not count N – the frequency of occurrence of the pattern word in all 200 pages. When calculating retrieval $r(n)$, we suppose that $N = 160$ because there are a maximum of 159 correct words in the first 500 of the sequence (24) (consisting of 4982 elements) and it is quite possible for the number of correctly separated words ꙗко and ꙗ́ко to be about 160 (see also Table 3).

The experiments used some combination of parameters: $\alpha = 0.05, 0.1, 0.2$, $\beta = 0.005, 0.001$, $\tau = 15, 19$ and $\rho = \rho_1, \rho_2$. The results presented on Figures 21 and 22 show that the search process is successful and in the retrieval interval $[0, 0.6]$ for almost all used parameters of GHD. We notice that the best $r_1 = 0.519$ is achieved with s-distance when the parameters are $\alpha = 0.1$, $\beta = 0.01$, $\tau = 19$, $\rho = \rho_1$.

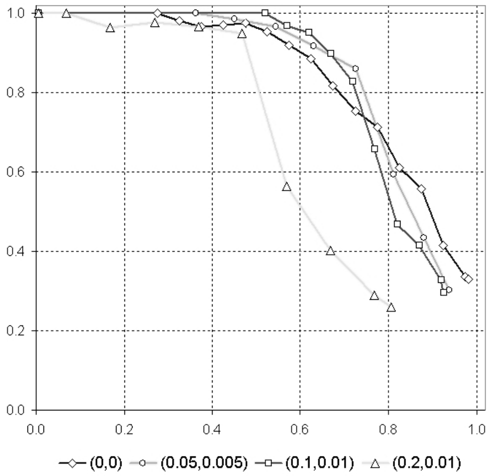


Fig. 21. **яко**: $(\alpha, \beta, 19, \rho_1)$,s

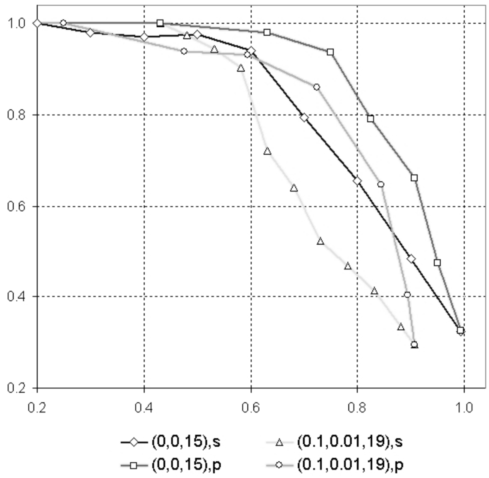


Fig. 22. **яко**: $(\alpha, \beta, \tau, \rho_2)$

The picture of LDM for two pairs of the pattern word (in the middle) and another word with approximately the same length can be seen on Fig. 23. The gray level corresponds to the distance from a current point to the set A or set B. If the left hand word is W_{i_1}, the pattern word is W_0 and the right hand word is W_{i_2}, then it is obvious that $H(W_0, W_{i_1}) > H(W_0, W_{i_2})$, i.e. $i_1 > i_2$ as word indexes in the sequence (23).

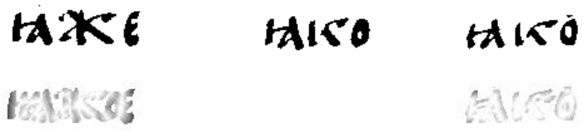


Fig. 23. **яко**: LDM

The Table 3 presents numerical data for an experiment with parameters

n	50	100	150	200	250	300	350	400	450	500
Correct $m(n)$	49	92	113	129	139	144	147	151	155	157
$m(n) - m(n-50)$	49	43	19	13	10	5	2	4	4	2
Retrieval $r(n)$	0.31	0.57	0.71	0.81	0.87	0.90	0.92	0.94	0.97	0.98
Precision $p(n)$	0.98	0.92	0.75	0.65	0.56	0.48	0.42	0.38	0.34	0.31

Table 3. The word **яко**: $(0, 0, 15, \rho_1)$,s

$(0, 0, 15, \rho_1)$, s-distance, and gives us the confidence for the practical usefulness of our methods and software even in handwritten texts.

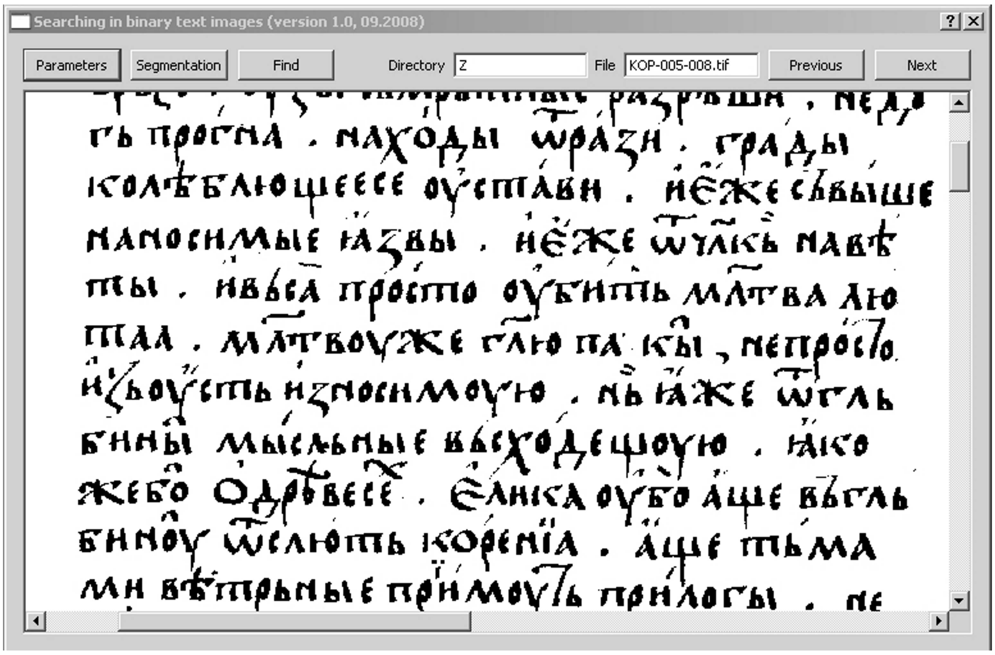
Other similar experiments with various documents in several languages and period can be found in Kirov [13].

7. Software. The following is a brief overview of the most important parts of a software system for searching in binary text images and the main steps in the searching process.

The input data are a collection of files representing a text document. Each file is one page image. Many graphic formats are acceptable: TIF, JPG, PGN, GIF, etc. We suppose that the input images are 1 Bpp (black and white) and only step for removing noise is applied. Precise binarization and improving text quality are not important in our approach because we compare images and there is no significant difference between comparing noisy or less noisy images.

Three main steps are essential for successful word searching: segmentation, searching and result representation. Segmentation is an important step for word searching. Only correctly segmented words have a chance of being included in list (23). Wa apply lines determination using simple histogram of black pixels (horizontal projection) which is a relatively easy step in processing document images. If the lines are horizontal straight lines, the histogram has near zero values between lines. Small line slopes does not change the result (see [12]). To segment the words in a line, we use vertical projection – a histogram obtained by counting the number of black pixels in each vertical scan at a given horizontal position. If the words are well separated, the histogram should have areas of near zero values between words. Because the intervals between words are larger than between characters, it is easier to separate words than characters. Segmentation of words and characters is also an important step in every OCR process. As a result, every word is associated with a word image – a minimal rectangular frame that contains the corresponding word. So we consider any word image as a rectangle, which consists of white and black pixels. The black pixels form a set, which is used in calculating word similarities.

After page segmentation, we choose a pattern word image W_0 – this is a word for searching in the document pages. At the search step we measure the similarity (using GHD) of a segmented word W and the pattern word W_0 and then create the list (23). Before that we pose the word image W at a suitable position with respect to the pattern image W_0 simply calculating a translation vector with ends $w_0 \in W_0$ and $w \in W$. There are three options for defining the points w_0 and w: taking geometric centers, mass centers or the left sides of word images W_0 and W.



Directory Z

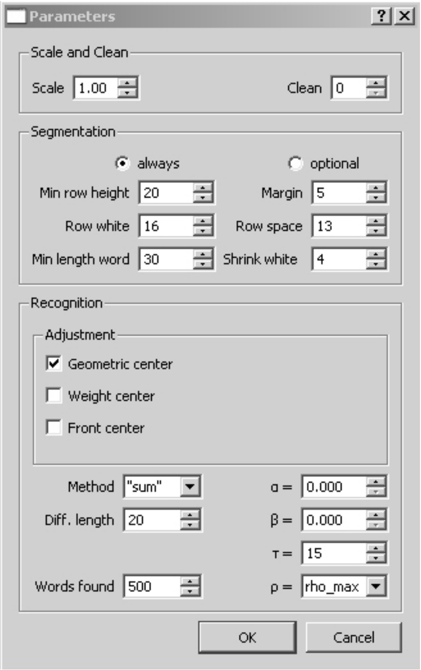
File KOP-005-008.tif

Fig. 24. The main window

The software system supports three user windows – main, `Parameters` and `Found words`. The main window (Fig. 24) is titled “Searching in binary text images” and presents to us a toolbar with control buttons. Also the current page of the document is displayed in this window. The names of the current directory and current file are placed on the toolbar. It is possible to go forward (`Next`) and backward (`Previous`) through the document pages. the `Parameters` button opens a new window – `Parameters` window; the `Segmentation` button starts segmentation step of the current page and the `Find` button starts searching. It activates the process of inspecting all pages for segmentation and measuring similarities of the segmented words and the pattern word.

The `Parameters` window (Fig. 25) allows the user to set various values to the control parameters for segmentation step and for search step. The segmentation parameters are described in [13]. In the `Recognition` frame we can choose the order in the list (23) to be based on p- or s-distance. The `Diff. length` parameter specifies the maximum length difference (in pixels) of pattern word W_0 and a given word W. The right hand column in this frame defines GHD parameters α, β, τ and ρ.

The `Found` window displays the result of the searching – the list (23) of words ordered by a similarity measure. We can see only a part of the sequence –



Parameters

Scale and Clean

Scale 1.00 Clean 0

Segmentation

● always ○ optional

Min row height 20 Margin 5

Row white 16 Row space 13

Min length word 30 Shrink white 4

Recognition

Adjustment

☑ Geometric center

□ Weight center

□ Front center

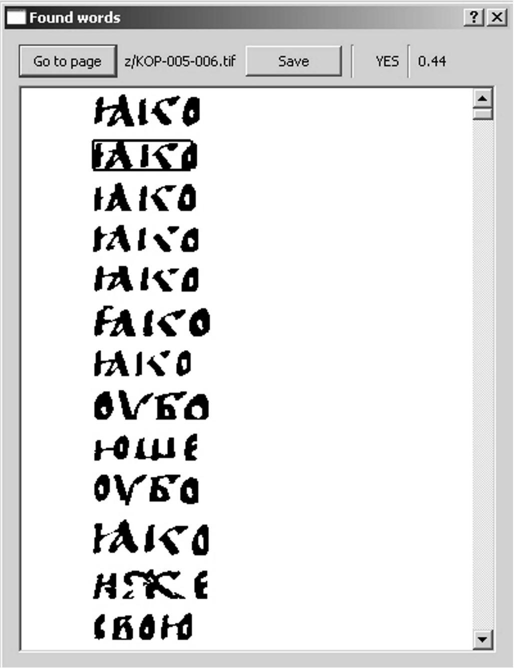
Method "sum" α = 0.000

Diff. length 20 β = 0.000

τ = 15

Words found 500 ρ = rho_max

Fig. 25. The `Parameters` window



Found words

z/KOP-005-006.tif YES 0.44

Fig. 26. The `Found` window

the number of words in this window is set in the `Parameter` window in the field `Words found`. Pushing the `GoTo` button, the page image containing the marked word is displayed in main window.

The program code is written in C++ with the help of Qt – a cross-platform application development framework [21].

8. Conclusions.

- The direct approach to searching words in binary text images could be applied successfully.
- HD (and its modifications) is a good choice for measuring word image similarities.
- GHD unify the HD approach; GHD contains many existing word matching methods and offers new methods by choosing various parameter sets.
- Obtaining a word sequence ordered by GHD for the given pattern word using primary and secondary sort keys corresponding to the s-distance and p-distance gives an additional advantage in practical profit.

- Creating an appropriate software tool for searching in binary text images gives us experimental power and shows that such a tool can be adapted as a completely functional user-oriented software product.
- The experiments with Bulgarian typewritten text, printed text and manuscript confirm the possibility of the wide application of our approach.
- Good perspectives for improvements exist both HD approach and software – for example searching a part of word, composing a pattern word from well separated letters, etc. (see [13]).

REFERENCES

[1] Andreev A., N. Kirov. Hausdorff Distance and Word Matching. In: Proc. Intern. Workshop Computer Science and Education, June 3-5, 2005, Borovetz-Sofia, Bulgaria, 19–28.

[2] Andreev A., N. Kirov. Some Variants of Hausdorff Distance for Word Matching. Review of the National Center for Digitization **12** (2008), 3–8.

[3] Andreev A., N. Kirov. Text Search in Document Images Based on Hausdorff Distance Measures. In: Proc. Intern. Conf. Computer Systems and Technologies (CompSysTech'08), June 12-13, 2008, Gabrovo, Bulgaria, II.5-1–II.5-6.

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

[5] Azencott R., F. Durbin, J. Paumord. Robust recognition of buildings in compressed aerial scenes. In: Proc. Intern. Conf. Image Proccessing, Lausanne, 1996.

[6] Baudrier E., F. Nicolier, G. Millon, Su Ruan. Binary-image comparison with local-dissimilarity quantification. Pattern Recognition **41** (2008), 1461–1478.

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

[8] Gatos B., T. Konidaris, K. Ntzios, I. Pratikakis, S. J. Perantonis. A Segmentation-free Approach for Keyword Search in Historical Typewritten Documents. In: Proc. Eight Int. Conf. on Document Analysis and Recognition (ICIDAR'05), 2005, 54–58.

[9] Huttenlocher D., G. Klanderman, W. Rucklidge. Comparing Images Using the Hausdorff Distance. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, **15** (1993), No 9, 850–863.

[10] Junker M., R. Hoch, A. Dengel. On the Evaluation of Document Analysis Components by Recall, Precision, and Accuracy. In: Proc. Fifth Intern. Conf. on Document Analysis and Recognition, (ICDAR'99), 20-22 Sep 1999, Bangelore, India, 1999, 713–716.

[11] Konidaris T., 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. *Intern. J. Document Analysis and Recognition*, **9** (2007), 167–177.

[12] Kirov N. Words Retrieval from Text Images. In: Proc. of the Fourth Intern. Workshop Computer Science and Education, June 6-8, 2008, Borovetz-Sofia, Bulgaria, 102–107.

[13] Kirov N. A Software Tool for Searching in Binary Text Images. Review of the National Center for Digitization, 2008.

[14] Lu Y., C. L. Tan, W. Huang, L. Fan. An Approach to Word Image Matching Based on Weighted Hausdorff Distance. In: Proc. of the Sixth Intern. Conf. Document Analysis and Recognition (ICDAR'01), 10-13 September 2001, Seattle, USA, 921–925.

[15] Paumard J. Robust comparison of binary images. *Pattern Recognition Letters*, **18** (1997), 1057–1063.

[16] Sim D.-G., O.-K. Kwon, R.-H. Park. Object Matching Algorithms Using Robust Hausdorff Distance Measures, *IEEE Trans. on Image Processing* **8** (1999), No 3, 425–429.

[17] Son H.-J., S.-H. Kim, Ji-Soo Kim. Text image matching without language model using a Hausdorff distance. *Information Processing & Management*, **44** (2008), Issue 3, 1189–1200.

[18] Дигитална Народна библиотека Србије, Ћирилски рукописи, Збирка словенских рукописа Јернеја Копитара, Зборник "Златоуст" [Digital National library of Serbia, Cyrillic manuscripts, Jernej Kopitar's collection of slavic manuscripts, Zbornik "Zlatoust"]
`http://www.digital.nbs.bg.ac.yu`, 1.12.2008

[19] Българска христоматия, съст. И. Вазов, К. Величков, ч.1 Проза, Книжарница на Д. В. Манчов, Пловдив, Свищов, Солун, 1884.

[20] Стоян Кендеров, Страници от музикалния живот в град Пазарджик, 2008.

[21] Trolltech: `http://trolltech.com`, 1.12.2008

Andrey Andreev
Institute of Mathematics and Informatics
Bulgarian Academy of Sciences
Acad. G. Bonchev Str., Bl. 8
1113 Sofia, Bulgaria
e-mail: `aandreev@math.bas.bg`

Nikolay Kirov
Department of Informatics
New Bulgarian University
21, Montevideo Str.
1618 Sofia, Bulgaria
e-mail: `nkirov@nbu.bg`
and
Institute of Mathematics and Informatics
Bulgarian Academy of Sciences
Acad. G. Bonchev Str., Bl. 8
1113 Sofia, Bulgaria
e-mail: `nkirov@math.bas.bg`
WEB: `http://www.math.bas.bg/~nkirov`

Received December 9, 2008
Final Accepted March 3, 2009