2009 10th International Conference on Document Analysis and Recognition

Word Image Matching Based on Hausdorff Distances*

Andrey Andreev and Nikolay Kirov
Institute of Mathematics and Informatics, BAS
"Acad. G. Bonchev" Str., Bl. 8, 1113 Sofia, Bulgaria
aandreev@math.bas.bg, nkirov@math.bas.bg

Abstract

Hausdorff distance (HD) and its modifications provides one of the best approaches for matching of binary images. This paper proposes a formalism generalizing almost all of these HD based methods. Numerical experiments for searching words in binary text images are carried out with old Bulgarian typewritten text, printed Bulgarian Chrestomathy from 1884 and Slavonic manuscript from 1574.

1. Introduction

Optical character recognition (OCR) is widely used approach for converting text images into text file. This step allows conducting text retrieval from scanned document images. OCR algorithm recognizes every character mapping it to a number, which is called code. Unfortunately often human efforts are needed to correct OCR errors which is quite tedious job. This is a consequence of bad original source or bad scanning process; old letters, outside the coding tables; old grammar; obsolete words, phrases and idioms; absence of 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 an efficient well-known task.

For word searching we suggest a different approach: words are searched in text images, obtained directly by scanning process (see [1], [2]) instead of applying OCR and searching in a text file. Organizing retrieval of words, similar to a given pattern word, by searching in the set of binary text images is an idea presented also in [4] and [10].

The main goals of this paper are:

• to propose a new method for estimating the similarity between two binary images in order to generalize and to unify the existing image matching methods based on Hausdorff distance;

• to check numerically the efficiency of generalized HD method when it is applied for word matching in typewritten, printed and handwritten historical documents.

*This work has been partially supported by Grant No. DO02-275/2008, Bulgarian NSF, Ministry of Education and Science.

2. Hausdorff distances for set similarities

The Hausdorff distance (HD) between two closed and bounded subsets A and B of a given metric space M is defined by

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

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

$$h(A,B) = \max_{a\in A} d(a,B), \;\; d(a,B) = \min_{b\in B} \rho(a,b). \qquad (2)$$

$d(a,B)$ is the distance from a point a to the set B, and $\rho(a,b)$ is a point distance in the metric space M.

HD looks very attractive for measuring the similarity between images as plane sets. Unfortunately, the HD (1) does not meet requirements of robustness. Many attempts have been 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 more points have to be included and in such way decreasing the influence of eventual presence of noise upon final evaluation of $H(A,B)$.

Let A and B be finite sets in the plane which consist of N_A and N_B points respectively and let ρ be the Euclidean distance in R^2.

D. P. Huttenlocher *et al.* [5] proposed Partial Hausdorff Distance (PHD) for comparing images containing a lot of 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:

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

978-0-7695-3725-2/09 $25.00 © 2009 IEEE
DOI 10.1109/ICDAR.2009.173

This definition of $K^{th}_{a\in A}$ differs from the original one in [5], where the rating order in (3) is in the opposite direction. The directed distance for PHD is

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

The idea of J. Paumard [8] is that we do not take into account the L closest neighbours 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 $\{\rho(a,b) : b \in B\} = \{\rho(a,b_i), i = 1,\dots,N_B\}$, i.e.

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

Now the directed Censored Hausdorff Distance (CHD) is defined by

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

For comparing two images obtained by adding randomly black and white dots to one of them the recommended values in [8] for the parameters are $K = 0.1N_A$ and $L = 0.01N_B$.

M.-P. Dubuisson and A. Jain [3] examined 24 distance measures of Hausdorff type to determine to what extend two finite sets A and B on the plane differ. Based on numerical behavior of these distances on synthetic images containing various levels of noise they introduced Modified Hausdorff Distance (MHD) with directed distance

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

In 1999 D.-G. Sim *et al.* [9] described two modifications 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

$$h_{\mathrm{M}}(A,B) = \frac{1}{N_A}\sum_{a\in A} f_\tau(d(a,B)), \tag{7}$$

where the function $f_\tau : R_+ \to R_+$ is increasing and has an unique minimum value at zero. They introduce one simple function with these properties

$$f_\tau(x) = \min\{x,\tau\}, \tag{8}$$

for a given $\tau > 0$. The recommended interval of τ is $[3,5]$ for their purposes.

The directed distance of LTS-HD is defined by

$$h_{\mathrm{LTS}}(A,B) = \frac{1}{N_A - K + 1}\sum_{i=K}^{N_A} d(a_i,B), \tag{9}$$

where $1 \le K \le N_A$ and $a_1, a_2, \dots, a_{N_A}$ are the points of A for which (3) is valid. The authors suggest $K/N_A = 0.2$ for comparing noisy binary images contaminated by Gaussian noise.

3. A new approach to HD similarity measures

Let us suppose there is a linear order of the points of the set $A = \{a_1, a_2, \dots, a_{N_A}\}$. For every $a_k \in A$ we calculate the distances from a_k to all points in B, as follows:

$$\begin{aligned}
&d_{k1} = \min_{b\in B}\rho(a_k,b) = \rho(a_k,b_{k1}),\\
&d_{k2} = \min_{b\in B\setminus\{b_{k1}\}}\rho(a_k,b) = \rho(a_k,b_{k2}),\\
&\dots,\\
&d_{kl} = \min_{b\in B\setminus\{b_{k1},\dots,b_{kl-1}\}}\rho(a_k,b) = \rho(a_k,b_{kl}),\\
&\dots
\end{aligned} \tag{10}$$

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

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

Let the matrix D be defined by

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

For a given $1 \le l \le N_B$, we define a new matrix D_l:

$$D_l = \left(d^{(l)}_{ij}\right), i = 1,\dots,N_A, j = 1,\dots,N_B$$

interchanging the rows of the matrix D so that the elements of l-th column are sorted, i.e. satisfying the following inequalities:

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

Let $1 \le k \le N_A$ and $1 \le l \le N_B$ be integer numbers. We define two Generalized Hausdorff Distances (GHD) using the following directed distances:

$$h^{(p)}_{k,l}(A,B) = d^l_{kl} \tag{11}$$

and

$$h^{(s)}_{k,l}(A,B) = \frac{1}{N_A - k + 1}\sum_{i=k}^{N_A} d^l_{il}. \tag{12}$$

We denote (11) by p-GHD and (12) by s-GHD. These definitions generalize all Hausdorff based distances mentioned above, which can be represented by their directed distances as follows:

HD (2): $h(A,B) = h^{(p)}_{1,1}(A,B) = d^1_{11}$;

PHD (4): $h_K(A,B) = h^{(p)}_{K,1}(A,B) = d^1_{K1}$;

CHD (5): $h_{K,L}(A,B) = h^{(p)}_{K,L}(A,B) = d^L_{KL}$;

MHD (6): $h_{\mathrm{MHD}}(A,B) = h^{(s)}_{1,1}(A,B) = \dfrac{1}{N_A}\sum_{i=1}^{N_A} d^1_{i1}$;

LTS-HD (9): $h_{\mathrm{LTS}}(A,B) = h^{(s)}_{K,1}(A,B)$.

We parameterize GHD replacing k and l in (11) and (12) by parameters α and β:

$$\alpha = \frac{k-1}{N_A}, \quad \beta = \frac{l-1}{N_B}. \tag{13}$$

Since $1 \le k \le N_A$ and $1 \le l \le N_B$ we have $\alpha, \beta \in [0,1)$.

In practice of image comparison, we have upper bounds for the distances between the points of any two images. Thus we define bounded modifications of point distances:

$$\rho^{(\tau)}(a,b) = \min\{\rho(a,b), \tau\}, \tag{14}$$

where τ is a positive number and $\rho(a,b)$ can be any point distance. The three most frequently used ones are Euclidean – $\rho_2(a,b) = \sqrt{(a_1-b_1)^2 + (a_2-b_2)^2}$, Manhattan – $\rho_1(a,b) = |a_1-b_1| + |a_2-b_2|$ and Chebyshev – $\rho_\infty(a,b) = \max\{|a_1-b_1|, |a_2-b_2|\}$, where $a = (a_1,a_2)$, $b = (b_1,b_2)$. Replacing ρ with $\rho^{(\tau)}$ in formulas (10) we introduce a new parameter τ for GHD. So for defining a concrete p- or s-GHD, we have to choose values for the parameters α, β, ρ and τ. Note that M-HD (7) with the function (8) coincides with MHD (6) applying $\rho^{(\tau)}$ for point distance.

3.1. Measuring searching effectiveness

The effectiveness of searching methods is usually given by standard estimations of recall and precision (see M. Junker *et al.* [6]). 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:

$$\{W_i\}_{i=0,1,\ldots} \tag{15}$$

which is ordered according to a similarity measure H, i.e. $H(W_i, W_0) \le H(W_j, W_0)$ for every $i < j$.

For a positive integer n, let $m(n) \le n$ be the number of words among the first n words of (15) that coincide with W_0 as words. Then recall $r(n)$ and precision $p(n)$ are defined by

$$r(n) = \frac{m(n)}{N} \quad \text{and} \quad p(n) = \frac{m(n)}{n}. \tag{16}$$

$m(n)$ is nondecreasing function and the graph of

$$P : D \subset [0,1] \to [0,1] \text{ defined by } P(r(n)) = p(n) \tag{17}$$

represents the effectiveness of searching methods. That sort of graphs are drawn on Figs 2, 3, 5 and 7.

4. Experiments

We define two implementations of s- and p-GHD denoted by:

– (α,β,τ),p – the sorting algorithm for producing the word sequence (15) uses primary sort key p-GHD and secondary sort key s-GHD. This approach avoids the discontinuity of p-GHD (see [1], and [2]) when the words in the sequence (15) are divided into a few classes, which correspond to equal distances to the pattern.

– (α,β,τ),s – the sorting algorithm uses primary sort key s-GHD and secondary sort key p-GHD.

In all experiments $n \in [1, 500]$.

4.1. Typewritten text

Bulgarian typewritten text of 333 bad quality pages (Fig. 1) is the data used in our experiments (see also [1] and [2]). A word **Пазарджик** is a pattern word W_0. It occurs 231

**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.**

Figure 1. Typewritten text

times in the text but the number of correct segmented words **Пазарджик** is 200, so we set $N = 200$. Figs 2 and 3 present graphs of the function (17) for this word. On Fig. 2 we can see that almost 80% of words **Пазарджик** are placed at the beginning of the sequence (15) in $(0.03, 0.005)$,s-case. The best precision 0.77 with maximum recall 0.95 is reached for $(0.03, 0.005)$,p. The remaining parameters for all cases are $\rho = \rho_2^{(\tau)}$ and $\tau = 15$.

The best results for the word **Пазарджик**, obtained in our experiments for s-case and $\rho = \rho_\infty^{(\tau)}$, are given in Fig. 3. We see that there is no best set of parameters – the maximum $r(n) = 0.825$ for $p(n) = 1$ is reached for $(0.01, 0.001)$ and $\tau = 15$ while for $r(n) \in [0.9, 0.975]$ the best parameters are $(0.03, 0.005)$ and $\tau = 19$.

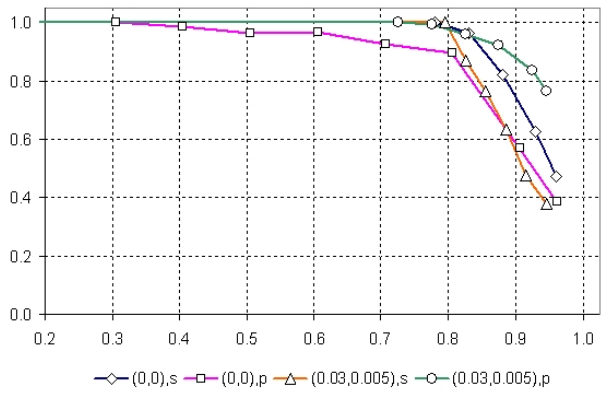

Figure 2.

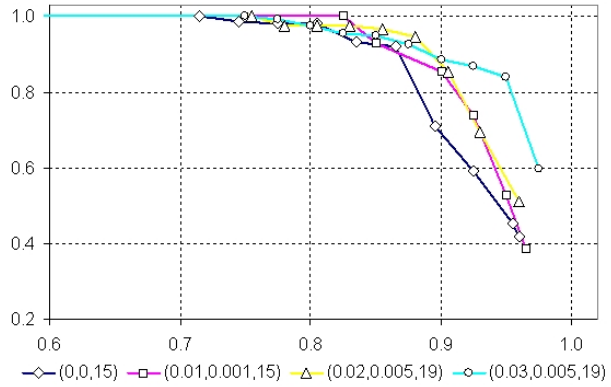

Figure 3.

4.2. Printed text

The carried out experiments are based on an old book (1884) – Bulgarian Chrestomathy, created by famous Bulgarian writers Ivan Vasov and Konstantin Velichkov (Fig. 4). Theoretically we can find all words in the printed text which coincide with a given pattern word under the assumption that scanned images are perfect. In this instance the quality of scanned images are quite bad. Many pages have slopes in the 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 reasons are the quality of images and the absence of OCR software because the text contains old and obsolete Bulgarian letters. Also spelling and grammar are quite different in modern Bulgarian language. For our experiments 200 images from about 1000 scanned pages are used. We choose a pattern word **всички**. It is tedious to count all words **всички** in all 200 pages, but we can estimate quite precisely their number. The best searching result give us 114 correct words in the first 500 of the sequence (15). The

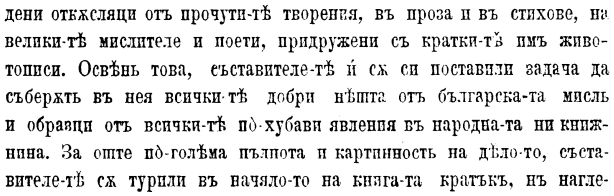
дени откѫсляци отъ прочути-тѣ творения, въ проза и въ стихове, на
велики-тѣ мислителе и поети, придружени съ кратки-тѣ имъ живо-
тописи. Освѣнь това, съставителе-тѣ и сѫ си поставили задача да
съберѫтъ въ нея всички-тѣ добри нѣщта отъ българска-та мисль
и образци отъ всички-тѣ по-хубави явления въ народна-та ни книж-
нина. За още по-голѣма пълнота и картинность на дѣло-то, съста-
вителе-тѣ сѫ турили въ начяло-то на книга-та кратъкъ, нъ нагле-

Figure 4. Printed text

total number of checked words with approximately same length is 7505 and the distribution of correct words is the reason for setting $N = 120$ and using this number in formulas (16).

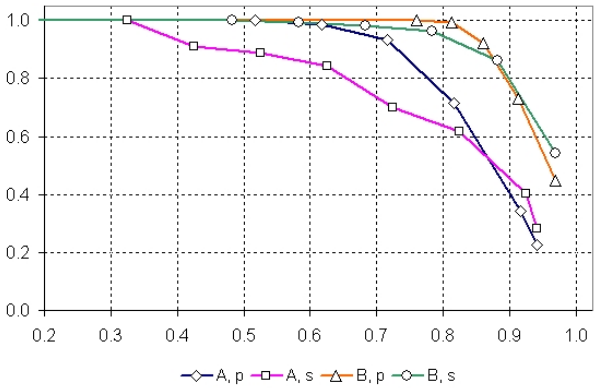

Figure 5.

Fig. 5 presents the results of applying GHD for $\alpha = 0.01$, $\beta = 0.001$, $\rho = \rho_2^{(\tau)}$ and $\tau = 15$. The graphics A,s and A,p are produced with the pattern word **всички**. – s- and p-case respectively.

In the text there are two cognate words **всичка** and **всичко**. When we count as correct all three of them, setting $N = 230$ the obtained results are better as it can be seen in Fig. 5, graphics B,p and B,s.

4.3. Handwritten text

The text under investigation is Slavonic manuscript collection (Fig. 6), "Zlatoust" (1574), 747 pages, but we consider 200 pages for the experiments. The segmentation is quite good due to the clerkly hand of the writer, and a relatively simple algorithm could separate rows and words. The pattern word is **ѵѧиѣѳ**. Occasionally the same word is written as **ѵѧиѣѳ**. We count both words as correct retrievals. There are two more words **ѵѧиѣѧ** and **ѵѧиѣь** which are very similar as images but have different meanings and we do not count them. When calculating $r(n)$, we suppose that $N = 160$ because there are maximum 159 correct words in the first 500 of the sequence (15), which consist of 4982

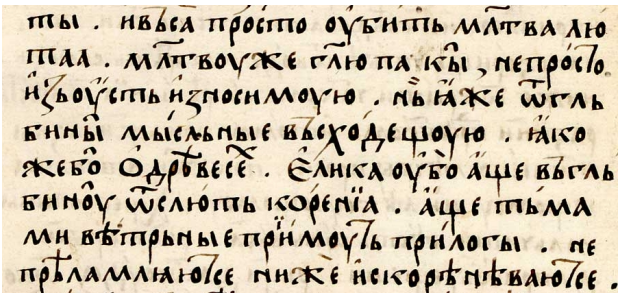

Figure 6. Handwritten text

words with approximately same length. The results pre-

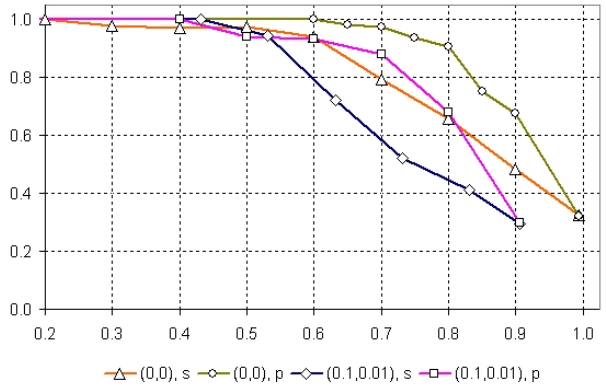

Figure 7.

sented on Fig. 7 show that the search process is the most successful for $\alpha = \beta = 0$ in p-case. The point distance is ρ_2, the parameter $\tau = 15$ for $\alpha = \beta = 0$ and $\tau = 19$ for $\alpha = 0.1$ and $\beta = 0.01$.

5. Conclusions

The experiments show that the direct approach for searching words in binary text images could be applied successfully in practice. HD and its modifications are a good choice for measuring word image similarities. GHD unifies the HD approach – GHD comprises of many existing word matching methods and offers new methods by choosing various values for the parameters α, β, τ and point distance, and processing s- or p-cases. The recommended values for α are in the interval $[0, 0.1]$ and for β in $[0, 0.01]$. All three distances ρ_2, ρ_1 and ρ_∞ can be used. The value of τ depends on image sizes but it must be greater than 5. There is no universal optimal parameter values for any scanned document and any searched word. The choice of good parameter values is made easier by using oriented software tool (see [7]). Quite acceptable results can be achieved for $\alpha = \beta = 0$ when the image quality is relatively good.

Obtaining a word sequence for a given pattern word ordered by p- and s-GHD, using primary and secondary sort keys, gives an additional advantage in practical aspects. The experiments with Bulgarian typewritten text, printed text and manuscript confirm the possibility of wide application of our approach.

References

[1] A. Andreev and N. Kirov. Hausdorff distance and word matching. In *Computer Science and Education*, pages 19–28, June 2005.

[2] A. Andreev and N. Kirov. Some variants of Hausdorff distance for word matching. *Review of the National Center for Digitization*, 12:3–8, 2008.

[3] M. P. Dubuisson and A. K. Jain. A modified Hausdorff distance for object matching. In *ICPR*, pages A:566–568, 1994.

[4] B. Gatos, T. Konidaris, K. Ntzios, I. E. Pratikakis, and S. J. Perantonis. A segmentation-free approach for keyword search in historical typewritten documents. In *ICDAR*, pages I: 54–58, 2005.

[5] D. Huttenlocher, D. Klanderman, and A. Rucklige. Comparing images using the Hausdorff distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(9):850–863, Sept. 1993.

[6] M. Junker, A. Dengel, and R. Hoch. On the evaluation of document analysis components by recall, precision, and accuracy. In *ICDAR*, pages 713–716, 1999.

[7] N. Kirov. A software tool for searching in binary text images. *Review of the National Center for Digitization*, 13:9–16, 2008.

[8] J. Paumard. Robust comparison of binary images. *Pattern Recognition Letters*, 18(10):1057–1063, Oct. 1997.

[9] D. G. Sim, O. K. Kwon, and R. H. Park. Object matching algorithms using robust Hausdorff distance measures. *IEEE Trans. Image Processing*, 8(3):425–429, Mar. 1999.

[10] H. J. Son, S.-H. Kim, and J. S. Kim. Text image matching without language model using a Hausdorff distance. *Inf. Process. Manage*, 44(3):1189–1200, 2008.