Hausdorff Distance and Word Matching

Andrey Andreev*, Nikolay Kirov**

Абстракт

An approach to word image matching based on Hausdorff distance is examined for bad quality typewritten Bulgarian text. A detailed computer experiments were carried out using 49 pages bad typed text. The results of several methods are compared including previously reported methods in the literature. The Hausdorff distance used in the paper differs slightly from ones used by other authors and the conclusion from the results is that our method outperforms them despite its simplicity.

Keywords: document text image, bitmap file, word matching, Hausdorff distance.

1 Introduction

Let A, B, $C, \ldots$ denote bounded sets on the plane and a, b, $c, \ldots$ be points on the plane with coordinates

$$a = (a_1, a_2),\ b = (b_1, b_2),\ c = (c_1, c_2),\ \ldots.$$

The Hausdorff distance (HD) between two bounded sets A and B is defined in [1] for the purposes of approximation of discontinues functions as

$$r_\alpha(A, B) = \max\{h_\alpha(A, B),\ h_\alpha(B, A)\}, \tag{1}$$

where

$$h_\alpha(A, B) = \max_{a \in A} \min_{b \in B} \rho(a, b), \tag{2}$$

$$\rho(a, b) = \max\left\{\frac{1}{\alpha}|a_1 - b_1|, |a_2 - b_2|\right\}, \tag{3}$$

and the parameter $\alpha > 0$. For $\alpha = 1$ we write

$$r(A, B) = r_1(A, B), \quad h(A, B) = h_1(A, B).$$

The parameter $\alpha \neq 1$ changed the equivalency of x and y axes. We accept that in typewritten text a word could be protracted in an arbitrary direction and therefore we set $\alpha = 1$.

*Institute of Mathematics and Informatics, Akad. G. Bonchev str., bl.8, Sofia, Bulgaria

**New Bulgarian University, Montevideo str. 21, Sofia, Bulgaria

This research has been supported by a Marie Curie Fellowship of the European Community programme "Knowledge Transfer for Digitalization of Cultural and Scientific Heritage in Bulgaria" under contract number MTKD-CT-2004-509754. The work has been done while the authors were at the National Center of Scientific Research "Demokritos", Athens, Greece.

In 1994 Dubuisson and Jain [2] examined 24 distance measures of Hausdorff type for determination to what extend two point sets on the plane A and B differ. In case when the sets A and B consist of N_A and N_B points along with (3) changed to Euclidean distance

$$\rho(a,b) = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2}, \tag{4}$$

they use

$$h(A,B) = \frac{1}{N_A} \sum_{a \in A} \min_{b \in B} \rho(a,b), \tag{5}$$

and claim that among all 24 "distances" examined by them the "distance" (1), (4), (5) called in [2] "Modified Hausdorff Distance" (MHD) suites in best way the problem for object matching.

Similar approach called "Weighted Hausdorff Distance" (WHD) is used in [3] for finding word image matching method in English and Chinese document images. In WHD method instead of (5)

$$h(A,B) = \frac{1}{N_A} \sum_{a \in A} \omega(a). \min_{b \in B} \rho(a,b), \quad \sum_{a \in A} \omega(a) = N_A, \tag{6}$$

is used, where the weight $\omega(a) \geq 0$ depends on the position of the point (pixel) a in a Chinese character.

Let us define two spaces of sets on the plane:

1. $\Re_F = \{A :$ the point set A consists of N_A points in the plane $\}$;

2. $\Re_{BC} = \{A : A$ is bounded and closed set in the plane $\}$.

The requirement (1) to be a metric either in $\Re_{BC}$ or at least in $\Re_F$ is natural and desirable. It is easy to prove that both distances HD and (1), (2), (4) are metrics in the space $\Re_{BC}$ and since $\Re_F \subset \Re_{BC}$ they are metrics also in $\Re_F$ as it is mentioned in [1] and [2]. On the other hand (2) is not a suitable part of HD for image matching although the later results show its applicability in some extend.

At the same time both "distances" MHD and WHD which are described in [2] and [3] as promising and superior over other "distances" are not metrics in $\Re_F$. Their advantage to HD lies in substitution of (2) by (5) or (6) which results in failure to satisfy a triangle inequality.

Based on the above analysis and on the results of the experiments we propose to simplify (5) using

$$h(A,B) = \sum_{a \in A} \min_{b \in B} \rho(a,b), \tag{7}$$

and call the distance (1), (3), (7) Sum (or Simple) Hausdorff Distance (SHD). SHD is still not a metric in $\Re_{BC}$ since if $A, B, C \in \Re_{BC}$:

1. $r(A,B) \geq 0$, and $r(A,B) = 0$ iff $A \equiv B$;

2. $r(A,B) = r(B,A)$;

3. $r(A,C) \leq r(A,C) + r(B,C)$ is not always fulfilled. For example let

$$\begin{aligned} A &= \{(0,0),(1,0),(1,1),(0,1)\}, \\ B &= \{(0,4),(1,4),(1,5),(0,5)\}, \\ C &= \{(2,2),(1,2),(2,3)\}. \end{aligned}$$

Then $r(A,B) = 14,\ r(A,C) = 6,\ r(B,C) = 7$ and the last inequality is not true.

The SHD satisfies

$$r(A,C) \leq r(A \cup B, C) \leq r(A,C) + r(B,C)$$

which can be written as

$$|r(A \cup B, C) - r(A,C)| \leq r(B,C). \tag{8}$$

If we consider A as an word image, C – as a template and B – as noise then (8) is an estimation of the growth of the distance. The proof of (8) is simple:

- if $r(A \cup B, C) = h(A \cup B, C)$ then

$$\begin{aligned} r(A \cup B, C) &= \sum_{d \in A \cup B} \min_{c \in C} \rho(d,c) = \sum_{d \in A \setminus B} \min_{c \in C} \rho(d,c) + \sum_{d \in B} \min_{c \in C} \rho(d,c) \\ &= h(A,C) + h(B,C) \leq r(A,C) + r(B,C); \end{aligned}$$

- if $r(A \cup B, C) = h(C, A \cup B)$ then

$$\begin{aligned} r(A \cup B, C) &= \sum_{c \in C} \min_{d \in A \cup B} \max\{|c_1 - d_1|, |c_2 - d_2|\} \\ \leq \sum_{c \in C} \min_{d \in A} \max\{|c_1 - d_1|, |c_2 - d_2|\} &+ \sum_{c \in C} \min_{d \in B} \max\{|c_1 - d_1|, |c_2 - d_2|\} \\ &= h(C,A) + h(C,B) \leq r(A,C) + r(B,C). \end{aligned}$$

2 Word matching using SHD versus MHD, HD and some other methods

For now on we shall use the following terminology:

- **word image** – a rectangular image which pixels have values 0 (white) or 1 (black);
- **word** – a subset of word image with pixel values 1.

2.1 Segmentation

The determination of the rows on a given page is an easy step in processing our documents. We use horizontal projection for row extraction. If the rows are horizontal straight lines, the histogram has zero values between rows. The same is when the rows have small slopes.

Vertical projection is a common method in word image segmentation. The vertical projection is the histogram obtained by counting the number of black pixels in each vertical scan at a given position. While the words are well separated, the histogram should have zero values between word images.

2.2 Mass or geometric center of a segmented word

The segmentation process produces detached word images with individual sizes and in general the identical words are with different (word image) sizes. For the recognition step we have to compare the images and they should have equal sizes.

Let X and Y be two rectangular word images. To enlarge X and Y in order to equalize their sizes two approaches are used – to coincide their geometric centers (gc) or to coincide their mass centers (mc). In both cases we determine the smallest rectangle Z, which contains both images. All pixels which belong to $Z\backslash Y$ or $Z\backslash X$ are set to white (zero) in the extended images.

2.3 Distances used in computer experiments

The following distances will be tested numerically for estimation of similarity between two **words** A and B:

1. L_1: $\mathrm{L}_1(A,B) = \sum_{a\in(A\backslash B)\cup(B\backslash A)} 1$;
2. HD: $\mathrm{HD}(A,B) = r(A,B)$, where $r(A,B)$ is defined by (1), (2), (3);
3. HD_1: $\mathrm{HD}_1(A,B) = r(A,B)$, where $r(A,B)$ is defined by (1), (7) and
$$\rho(a,b) = \begin{cases} 0, & \text{if } a = b, \\ 1, & \text{otherwise.} \end{cases}$$
4. MHD: $\mathrm{MHD}(A,B) = r(A,B)$, where $r(A,B)$ is defined by (1), (3), (5);
5. SHD: $\mathrm{SHD}(A,B) = r(A,B)$, where $r(A,B)$ is defined by (1), (3), (7).

Using the distances defined above we carry out a series of computer word matching experiments. Real Bulgarian documents of typewritten text of 49 pages of bad quality as shown on Fig. 1 are the material from which a specified word is located and extracted. As templates serve the word images of three words ("така", "песни", "Пазарджик") with

както българите са се възхищавали от хубавите мелодии на
маанета, пластични кючеци и други песни, така и турците
са се любували на кръшните български хора и мелодични
народни песни.

Не малко музиканти са били турски цигани и са
свирили по български сватби, хорища, сборове и пр.

Фигура 1: Typewritten Bulgarian text

different number of letters – 4, 5 and 9 respectively. These templates are selected from the segmented words and are given on Fig. 2. The first word occurs 13 times, second one – 31 times and third one – 57 times.

Before using a given distance for estimation the difference between two images they must be adjusted with respect to either their geometric centers or to their mass centers. For example if SHD distance is applied combined with geometric center adjustment of images we denote this by SHD^{gc} otherwise we write SHD^{mc}. Measuring the effectiveness of

Фигура 2: Three template words

the distances (or methods connected with them) usually is given by standard estimations *Recall* and *Precision* [4]. Briefly, let us look for a word W in a collection of binary text images in which W occurs N times. Let the method Φ produce a sequence of words

$$\{W_i\}_{i=1,2,\ldots,} \tag{9}$$

ordered according to a specific for Φ criteria. For a given n ($n = 1, 2, \ldots$), let $n_1 \leq n$ be the number of words among the first n words of (9) that coincide with W. Note that n_1 is a function of n. Then we define

$$Recall_{\Phi}(n) = \frac{n_1}{N} \quad \text{and} \quad Precision_{\Phi}(n) = \frac{n_1}{n}, \tag{10}$$

as functions of n.

2.4 Experimental results

For the words "така", "песни" and "Пазарджик" computer results are summarized as:

1. Word "така": occurrence 13 times.

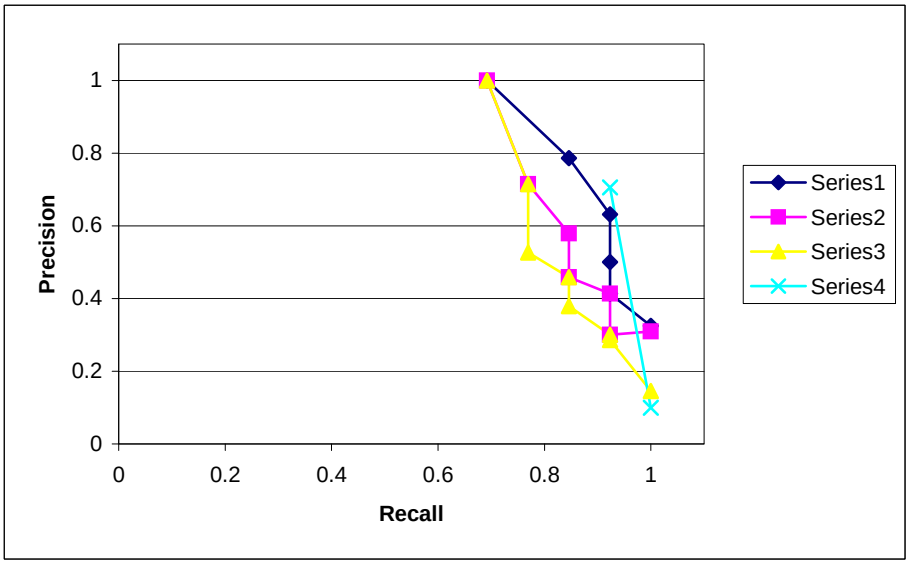
Series1: SHD^{gc}			
n	n_1	*Recall*	*Precision*
9	9	0.69	1.00
14	11	0.85	0.79
19	12	0.92	0.63
24	12	0.92	0.50
29	12	0.92	0.41
40	13	1.00	0.32

Series3: L_1^{gc}			
n	n_1	*Recall*	*Precision*
9	9	0.69	1.00
14	10	0.77	0.79
19	10	0.77	0.53
24	11	0.85	0.46
29	11	0.85	0.38
40	12	0.92	0.30
42	12	0.92	0.29
89	13	1.00	0.15

Series2: MHD^{gc}			
n	n_1	*Recall*	*Precision*
9	9	0.69	1.00
14	10	0.77	0.71
19	11	0.85	0.58
24	11	0.85	0.46
29	12	0.92	0.41
40	12	0.92	0.30
42	13	1.00	0.31

Series4: HD^{gc}			
n	n_1	*Recall*	*Precision*
17	12	0.92	0.71
131	13	1.00	0.10

The results from the tables above for "така" are plotted on Fig. 3



Фигура 3: Word "така"

2. Word "песни": occurrence 31 times.

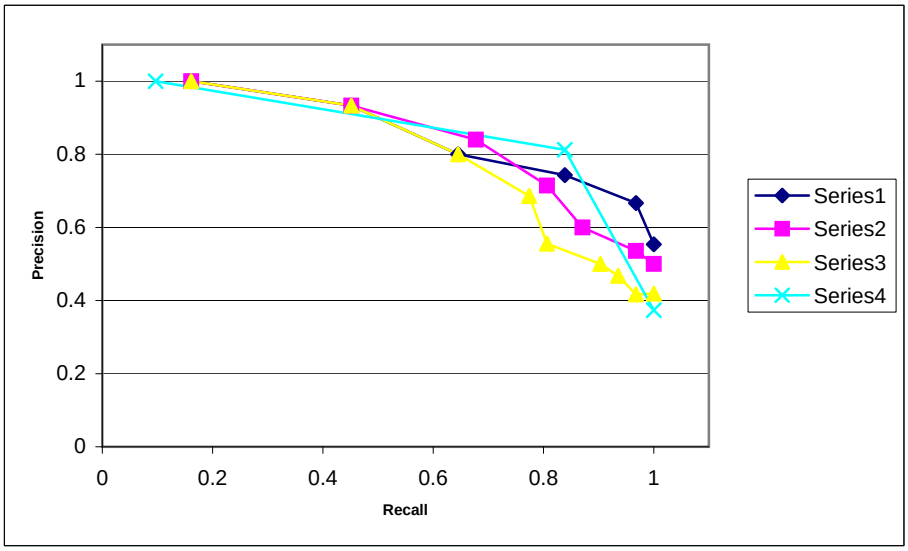
Series2: MHD^{gc}			
n	n_1	*Recall*	*Precision*
5	5	0.16	1.00
15	14	0.45	0.93
25	21	0.68	0.84
35	25	0.81	0.71
45	27	0.87	0.60
56	30	0.97	0.54
62	31	1.00	0.50

Series3: HD$_1^{gc}$			
n	n_1	*Recall*	*Precision*
5	5	0.16	1.00
15	14	0.45	0.93
25	20	0.65	0.80
35	24	0.77	0.69
45	25	0.81	0.56
56	28	0.90	0.59
62	29	0.94	0.47
72	30	0.97	0.42
74	31	1.00	0.42

Series1: SHD^{gc}			
n	n_1	*Recall*	*Precision*
5	5	0.16	1.00
15	14	0.45	0.93
25	20	0.64	0.80
35	26	0.84	0.74
45	30	0.97	0.67
56	31	1.00	0.55

Series4: HD^{gc}			
n	n_1	*Recall*	*Precision*
3	3	0.10	1.00
32	26	0.84	0.81
83	31	1.00	0.37

The results from the last 4 tables above are plotted on Fig. 4



Фигура 4: Word "песни"

3. Word "Пазарджик": occurrence 57 times.

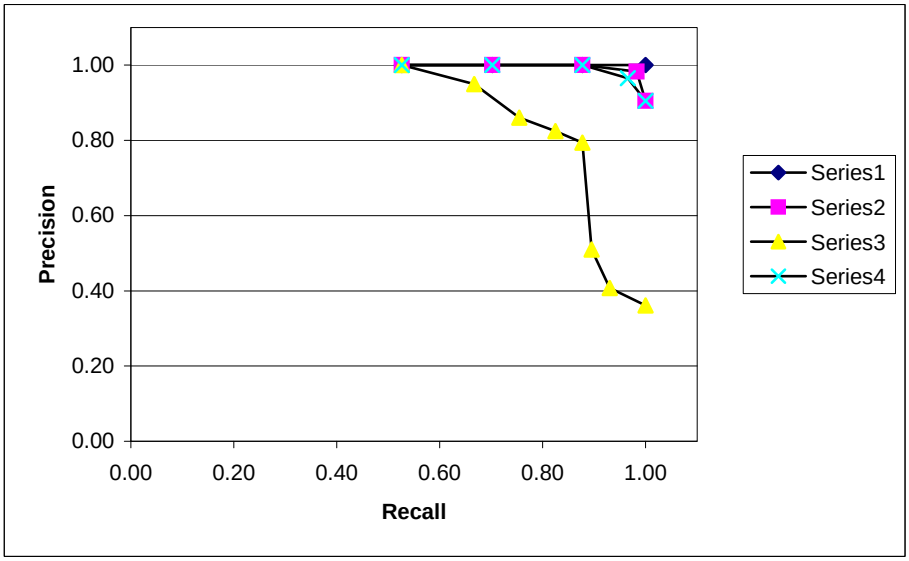
Series2: MHD^{gc}			
n	n_1	*Recall*	*Precision*
30	30	0.53	1.00
40	40	0.70	1.00
50	50	0.88	1.00
57	56	0.98	0.98
63	57	1.00	0.90

Series3: SHD^{wc}			
n	n_1	*Recall*	*Precision*
30	30	0.53	1.00
40	38	0.67	0.95
50	43	0.75	0.86
57	47	0.82	0.82
63	50	0.88	0.79
100	51	0.89	0.51
130	53	0.93	0.41
158	57	1.00	0.36

Series1: $\mathrm{SHD}^{gc} \equiv \mathrm{HD}^{gc}$			
n	n_1	*Recall*	*Precision*
30	30	0.53	1.00
40	40	0.70	1.00
50	50	0.88	1.00
57	57	1.00	1.00

Series4: L_1^{gc}			
n	n_1	*Recall*	*Precision*
30	30	0.53	1.00
40	40	0.70	1.00
50	50	0.88	1.00
57	55	0.96	0.96
63	57	1.00	0.90

The results for word "Пазарджик" are plotted on Fig. 5



Фигура 5: Word "Пазарджик"

4. Let us pose the following problem: how successfully can we find all words that begin with certain pattern of characters? For this aim we use the word "музика", Fig. 6, as 13 other (different from the word "музика") words in these 49 pages begin with the same characters.

Фигура 6: Word "музика"

These words occur 152 times and they are:

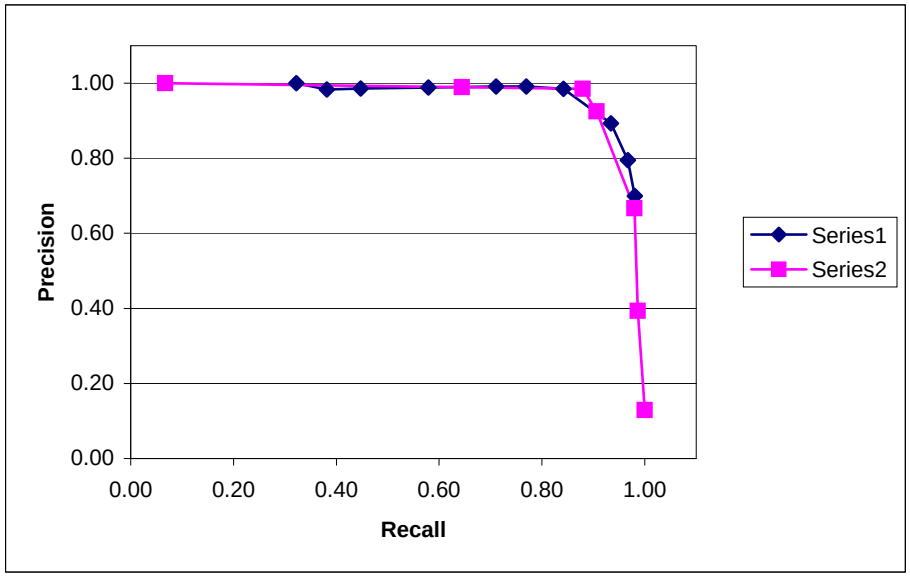
музиката	**музика**нт	**музика**нти	**музика**нтите
музиканта	**музика**нтски	**музика**лно	**музика**лни
музикална	**музика**лния	**музика**лната	**музика**лен
музикалните			

Let us note that because of comparing words of different length their horizontal adjustment is irrelevant while vertical adjustment is desirable. Computer results are given below.

Series1: SHD^{gc}			
n	n_1	*Recall*	*Precision*
49	49	0.32	1.00
59	58	0.38	0.98
69	68	0.45	0.99
89	88	0.58	0.99
109	108	0.71	0.99
118	117	0.77	0.99
130	128	0.84	0.99
159	142	0.93	0.89
185	147	0.96	0.80
213	149	0.98	0.70

Series2: HD^{gc}			
n	n_1	*Recall*	*Precision*
10	10	0.07	1.00
97	96	0.64	0.99
133	131	0.88	0.98
146	135	0.91	0.92
219	146	0.98	0.67
374	147	0.99	0.39
1157	149	1.00	0.13

The results for found words that begin with "музика" are plotted on Fig. 7



Фигура 7: Results for words that begin with "музика"

2.5 Conclusion

We process bad typewritten Bulgarian text for word matching using various distances. The results (plotted on Figs 3, 4, 5 and 7) show that:

- The general observation is that longer words are easier to be caught by all distances. This is expected because the longer word contains more specific information.
- The distance SHD^{gc} produces better results than other distances and therefore there is no need to complicate the definition of SHD (like MHD or WHD).
- Mass centered adjustment mc of word images is inappropriate for the purpose of word matching.

- Classic Hausdorff distance HD^{gc} does not loss ground to other approaches for such n for which

$$Recall(n) \leq 0.85.$$

  For words which contain more letters like "Пазарджик" the distance HD^{gc} works as good as "the best" method SHD^{gc}.

- L_1^{gc} distance produces the worst results. HD_1^{gc} method which is a sort of a combination of L_1^{gc} and SHD^{gc} behaves better, but evidently falls back to SHD^{gc}.

- The distance MHD^{gc} (originally given in [2] by (4), now changed to (3)) is slightly worse than SHD^{gc}.

- The measurement done by HD^{gc} distance could be considered as a "discontinuity". This explains the deterioration of the results produced by HD^{gc} for values of $Recall(n) \approx 1$. For example (see Fig. 3), for the short word "така" with occurrence 13 times HD^{gc} finds:

HD^{gc} distance	No. of words found n	No. of correct words n_1
3	17	12
4	114	1

  In this sense the other methods use practically continuous scale for ordering the spotted words.

Литература

[1] Bl. Sendov, Hausdorff Approximations, Kluwer Academic Publishers, 1990.

[2] 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.

[3] Y. Lue, C. L. Tan, W. Huang, L. Fan, *An Approach to Word Image Matching Based on Weighted Hausdorff Distance*, "6th ICDAR", 10-13 Sept. 2001, Seattle, USA.

[4] M. Junker, R. Hoch, A. Dengel, *On the Evaluation of Document Analysis Components by Recall, Precision, and Accuracy*, Proceedings ICDAR 99, Fifth Intl. Conference on Document Analysis and Recognition, Bangelore, India, 1999.