Introduction

Let A and B denote bounded sets on the plane and a and b be points on the plane with coordinates

$$a = (a_1, a_2),\ b = (b_1, b_2).$$

The Hausdorff distance (HD) between two bounded sets A and B is defined as

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

where

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

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

Dubuisson and Jain (1994) examined 24 distance measures of Hausdorff type for determination to what extend two point sets on the plane A and B differ. Let the sets A and B consist of N_A and N_B points and

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

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 "Modified Hausdorff Distance" (MHD) suites in best way the problem for object matching.

Similar approach called "Weighted Hausdorff Distance" (WHD) is used by Lue, Tan, Huang and Fan (2001) 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.

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).

Word matching using SHD versus other methods

- **word image** – a rectangular image which pixels have values 0 (white) or 1 (black);

- **word** – a subset of word image with pixel values 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.

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 or to coincide their mass centers. 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.

Distances used in computer experiments

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

1. $\mathsf{L}_1(A,B) = \sum_{a \in (A \setminus B) \cup (B \setminus A)} 1;$

2. $\mathsf{HD}(A,B) = r(A,B)$, where $r(A,B)$ is defined by (1), (2), (3);

3. $\mathsf{HD}_1(A,B) = r(A,B)$, where $r(A,B)$ is defined by (1), (7) and

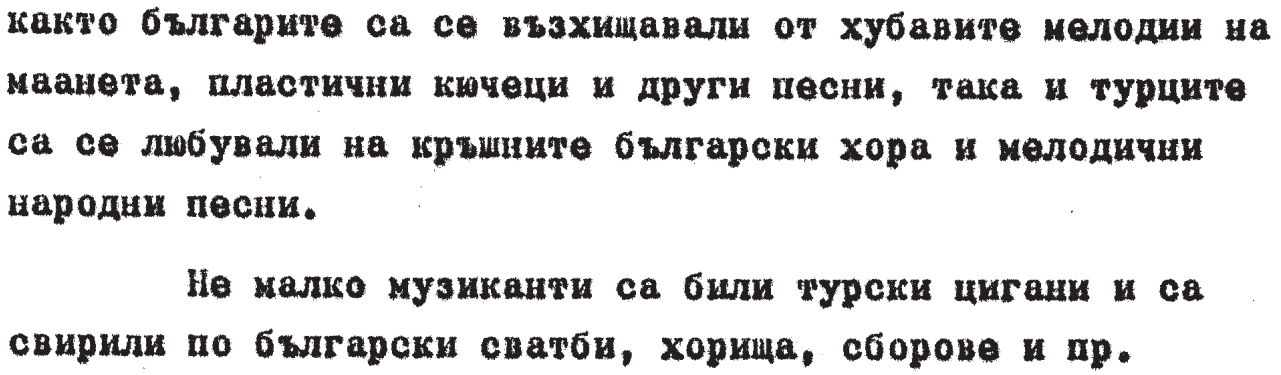
$$\rho(a,b) = \begin{cases} 0, & \text{if } a = b, \\ 1, & \text{otherwise.} \end{cases}$$

4. $\mathsf{MHD}(A,B) = r(A,B)$, where $r(A,B)$ is defined by (1), (3), (5);

5. $\mathsf{SHD}(A,B) = r(A,B)$, where $r(A,B)$ is defined by (1), (3), (7).

Computer word matching experiments

Real Bulgarian documents of typewritten text of 49 pages of bad quality are the material from which a specified word is located and extracted.



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.

As templates serve the word images of three words (“така”, “песни”, “Пазарджик”) with different number of letters – 4, 5 and 9 respectively.

така песни
Пазарджик

Three template words

The word “така” occurs 13 times,

“песни” – 31 times and

“Пазарджик” – 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* (gc) or to their *mass centers* (mc).

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 the distances

The effectiveness of the distances usually is given by standard estimations *Recall* and *Precision* – Junker, Hoch, Dengel (1999). Let us look for a word W in a collection of binary text images in which W occurs N times. Let a method produce a sequence of words

$$\{W_i\}_{i=1,2,\ldots,} \tag{8}$$

ordered according to a specific 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 (8) that coincide with W. Note that n_1 is a function of n. Then we define

$$\mathit{Recall}(n) = \frac{n_1}{N} \quad \text{and} \quad \mathit{Precision}(n) = \frac{n_1}{n}, \tag{9}$$

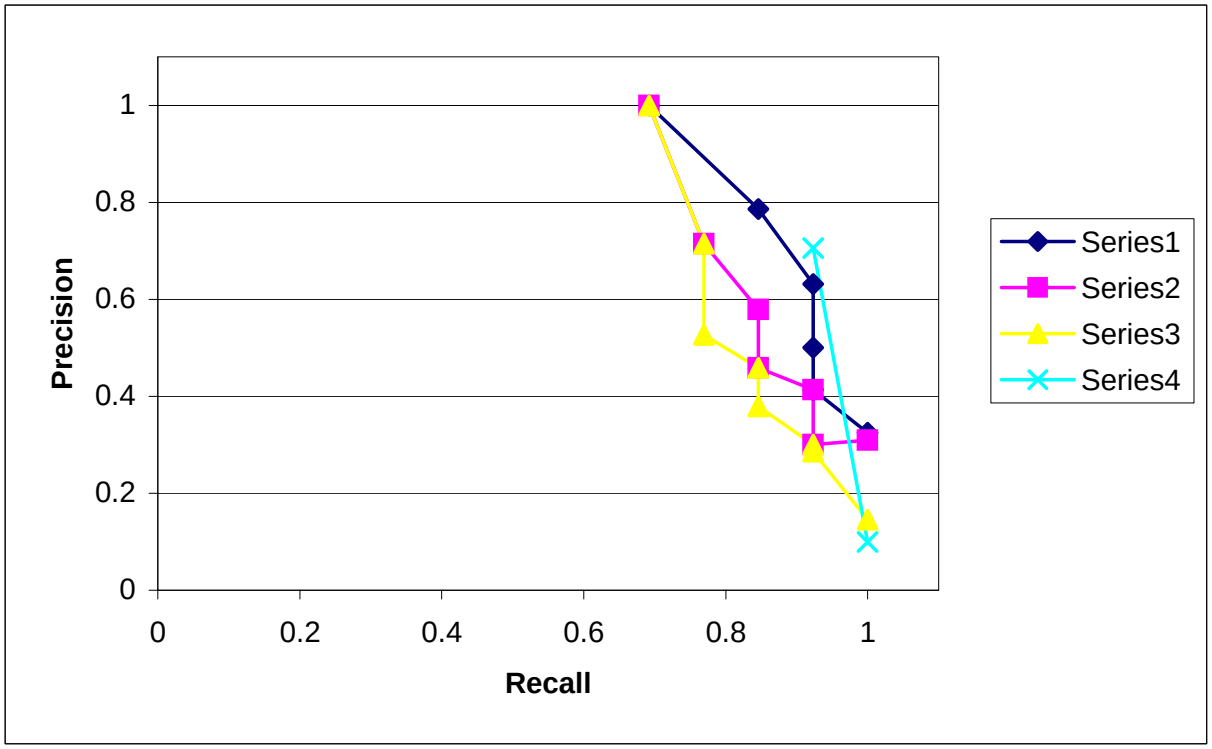
as functions of n.

Experimental results

Word "таka": occurrence 13 times.

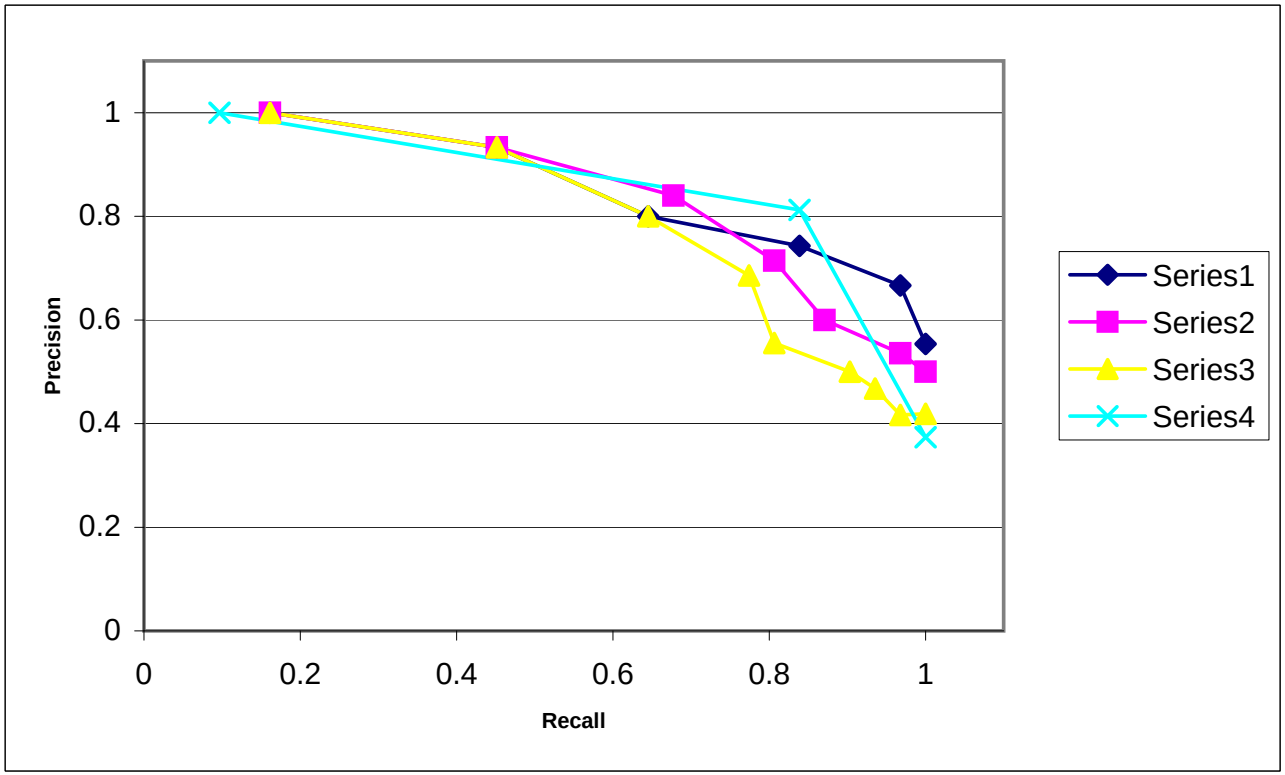
Series 1: SHD^{gc}			
n	n_1	*Recall*	*Precision*
9	9	0.69	1.00
14	11	0.85	0.79
19	12	0.92	0.63
40	13	1.00	0.32

Series 3: L_1^{gc}			
n	n_1	*Recall*	*Precision*
9	9	0.69	1.00
14	10	0.77	0.79
24	11	0.85	0.46
40	12	0.92	0.30
89	13	1.00	0.15



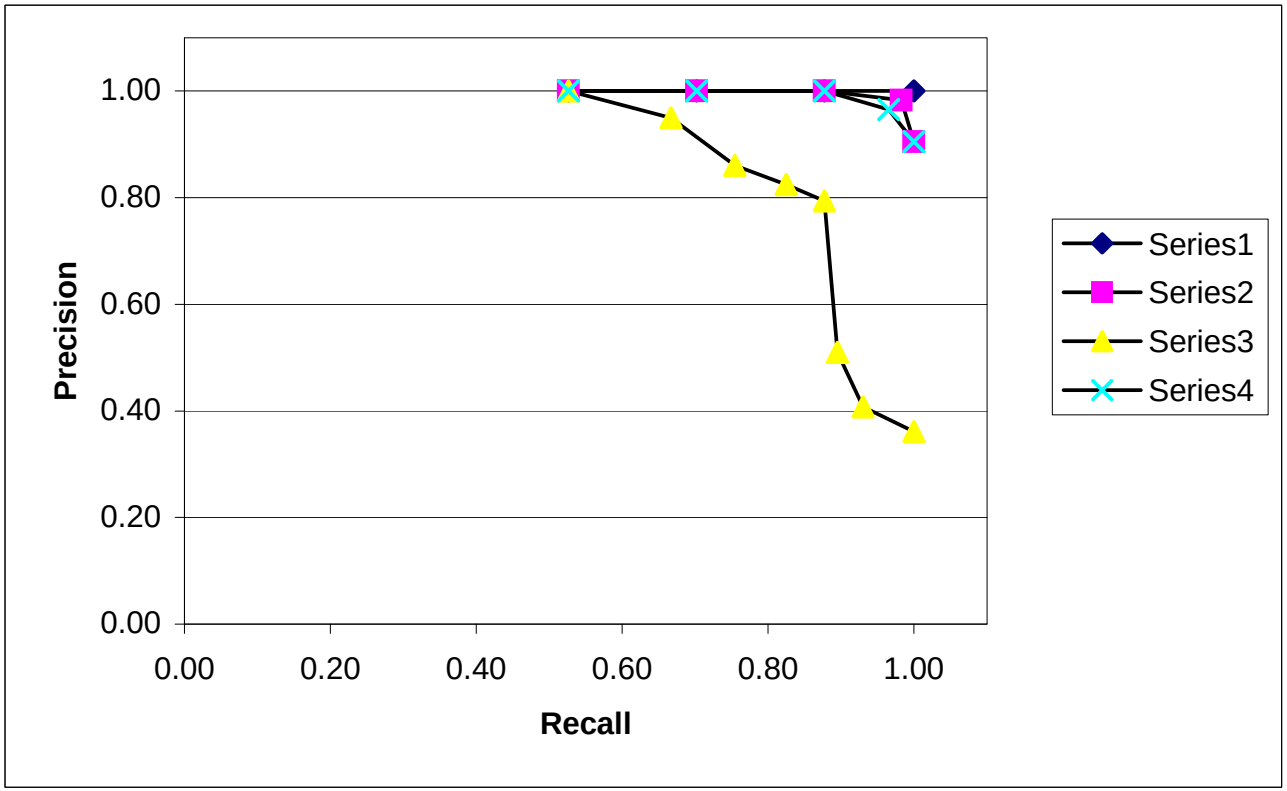
Series 1: SHD^{gc},
Series 2: MHD^{gc},
Series 3: L_1^{gc},
Series 4: HD^{gc}

Word “песни”: occurrence 31 times.



Series 1: SHD^{gc}, Series 2: MHD^{gc}, Series 3: HD_1^{gc}, Series 4: HD^{gc}

Word “Пазарджик”: occurrence 57 times.



Series 1: $\mathrm{SHD}^{gc} \equiv \mathrm{HD}^{gc}$, Series 2: MHD^{gc}, Series 3: SHD^{wc}, Series 4: L_1^{gc}

How successfully can we find all words that begin with certain pattern of characters? For this aim we use the word “музика”,

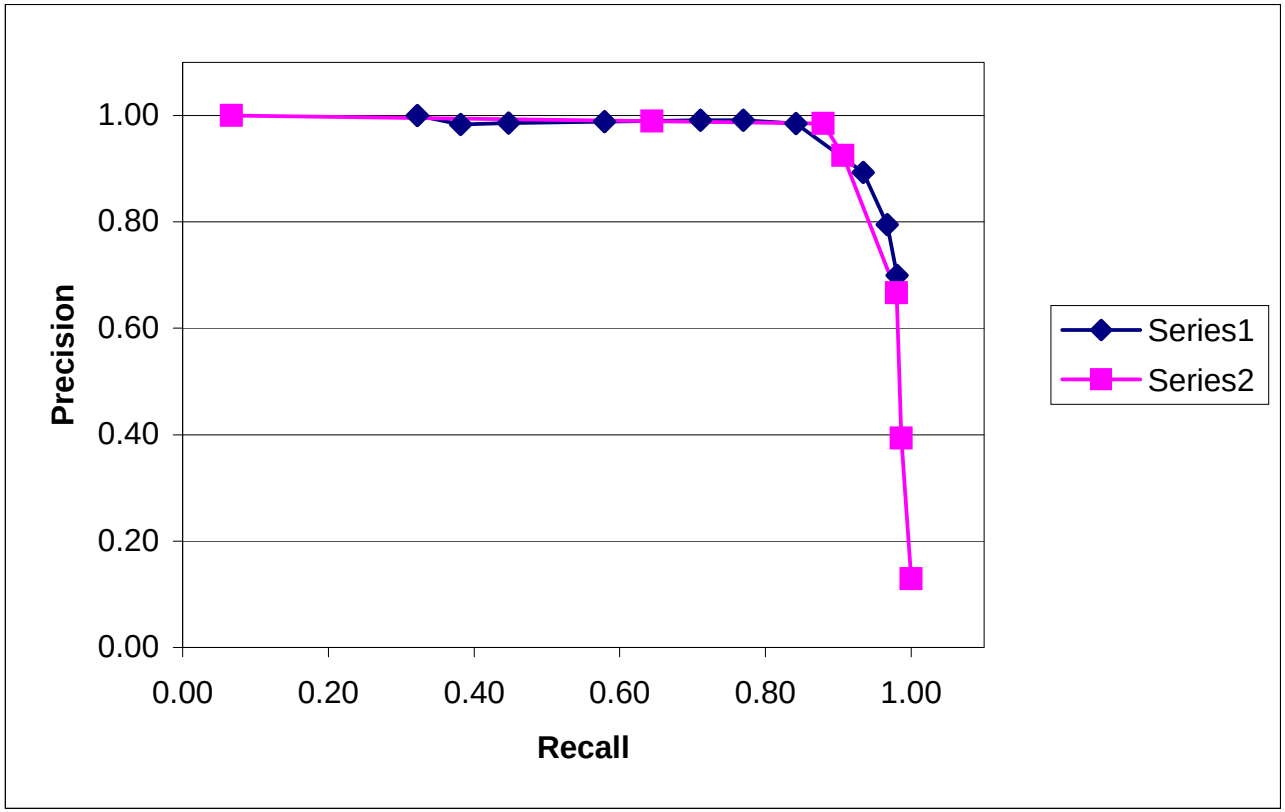
музика

as 13 other (different from the word “музика”) words in these 49 pages begin with the same characters. 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.

The results for found words that begin with “музика” are plotted on the following picture:



Series 1: SHD^{gc}, Series 2: HD^{gc}

Conclusion

We process bad typewritten Bulgarian text for word matching using various distances. The results 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

$$\mathit{Recall}(n) \leq 0.85.$$

  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 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, for the short word "таka" 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.