The third SEEDI International Conference:
***Digitization of cultural and scientific heritage***
September 13-15, 2007, Cetinje, Montenegro

# Methods for Word Image Matching

Andrey Andreev[1], Nikolay Kirov[2]

[1]Institute of Mathematics and Informatics, BAS, Sofia

[2]New Bulgarian University & Institute of Mathematics and Informatics, BAS, Sofia

This research has been supported by a Marie Curie Fellowship of the EC programme "Knowledge Transfer for Digitization of Cultural and Scientific Heritage in Bulgaria".

# Introduction

An approach to word image matching based on Hausdorff distance is examined for low quality typewritten documents in Bulgarian language. Computer experiments were carried out using 54 pages typewritten text. The results of several methods are compared. The goal is to find out which modification of Hausdorff distance suits satisfactory to the problem for word matching.

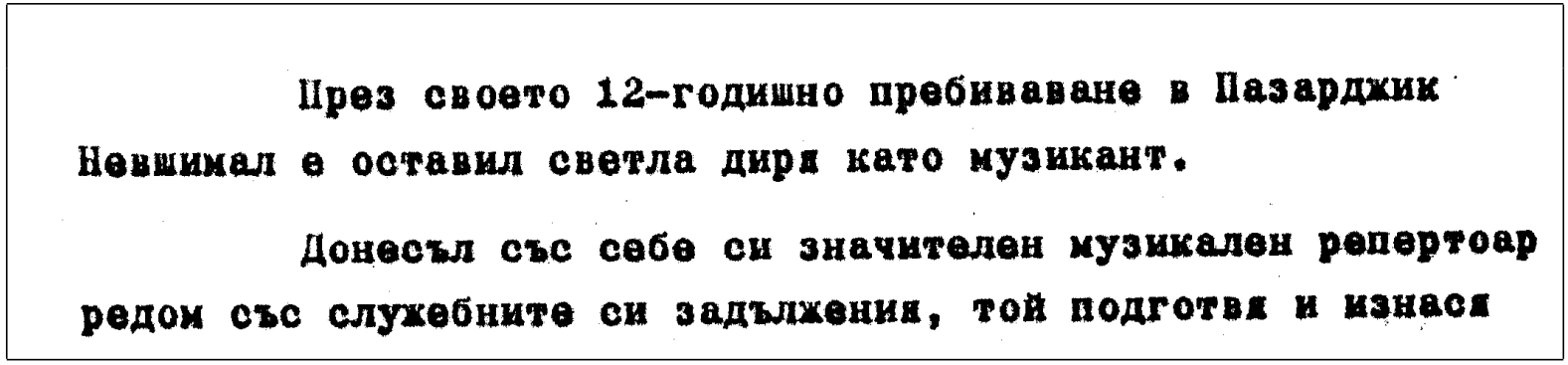
**Note. Why have we changed the title of our presentation?**

E. Baudrier, F. Nicolier, G. Millon, Su Ruan, *Binary-image comparison with local-dissimilarity quantification*, (Accepted in Pattern Recognition, July, 2007)

Taking into account the conclusions from the paper we decided change the topic and the title...

# The Problem

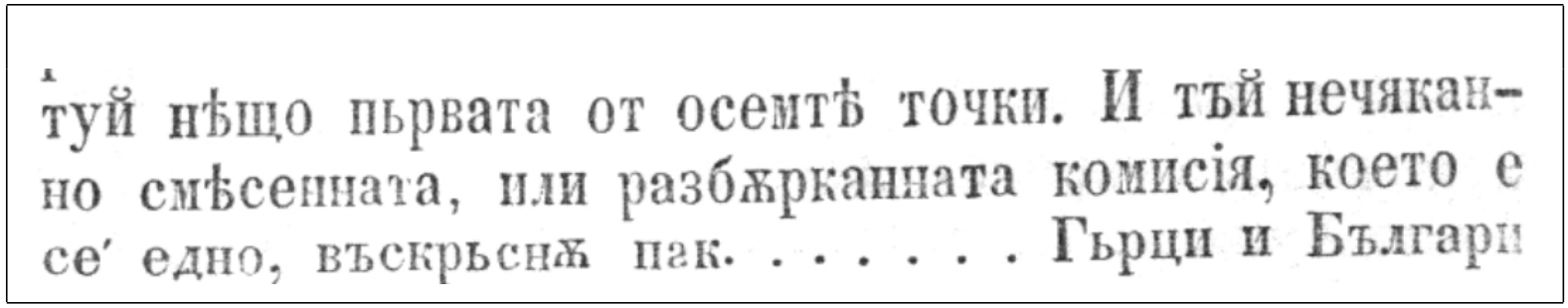
- We have a lot of pages of low quality typewritten documents.
- We have binary images of these pages (scanned).



През своето 12-годишно пребиваване в Пазарджик
Невшимал е оставил светла дира като музикант.

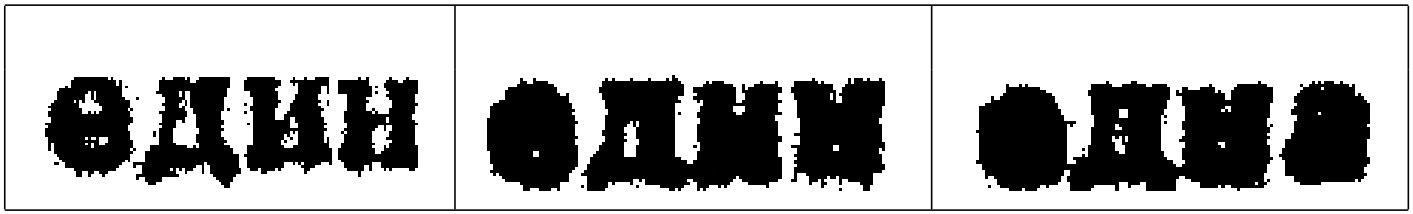
Донесъл със себе си значителен музикален репертоар
редом със служебните си задължения, той подготвя и изнася

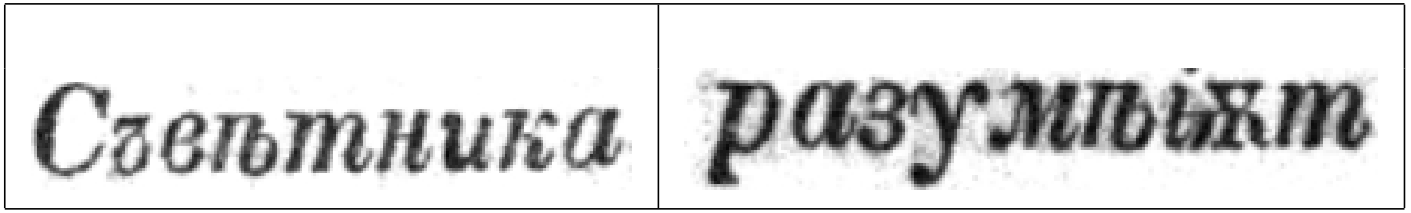
- It is difficult to OCR because of low quality, old grammar and spelling and old (not used now) letters.



туй нѣщо първата от осемтѣ точки. И тъй нечякан-
но смѣсенната, или разбѫрканната комисія, което е
се' едно, въскрьснѫ пак. . . . . Гърци и Българи

- It is difficult even for a man much less for the computer to understand every letter or word in the text like this below.





# The proposal for solving the problem

- To write a user frendly software tool, which implements a method for searching a word or phrase in a set of pages in form of binary images – B. Gatos, I. Pratikakis and S. J. Perantonis (2004).
- The core of such tool is the method for word comparison and ordering words in respect to their distances to a chosen pattern word.
- The pattern word can be a synthetic keyword (T. Konidaris, B. Gatos, at all. 2007) or can be a real word chosen from a page (feedback).

# Original Hausdorf Distance

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 the sets $A$ and $B$ is defined as

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

where

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

is called **directed distance** from $A$ to $B$. Here $D(a, B)$ is the distance between the point $a$ and the set $B$ and $\rho(a, b)$ is an arbitrary distance between the points $a$ and $b$. The most natural is the Euclidean distance

$$\rho(a, b) = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2} \tag{3}$$

but in calculations it is often used maximum distance.

$$\rho(a, b) = \max\left\{|a_1 - b_1|, |a_2 - b_2|\right\}. \tag{4}$$

# Censored Hausdorff Distance

The idea of José Paumard (1997) is that we do not take into account the $p$ closest neighbours of $a$ in $B$. $p$ is chosen as $\alpha\%$ of number of points in $B$, i.e. $p = \alpha N_B$. And we compute $D_\alpha(a, B)$ with the $p+1$-st closest neighbour of $a$ in $B$. Note that $D_0(a, B) = D(a, B)$.

Let us consider a set of numbers $X = \{x_1, x_2, \ldots, x_N\}$ where $x_i \leq x_{i+1}$. For $\alpha \in [0, 1]$, he defines $N_\alpha = \alpha N$ and $Q_\alpha\{X\} = x_{N_\alpha}$. Censored Hausdorff Distance (CHD) is:

$$CHD_{\alpha,\beta}(A, B) = \max\{h_{\alpha,\beta}(A, B),\ h_{\alpha,\beta}(B, A)\},$$

$$h_{\alpha,\beta}(A, B) = Q_{1-\beta}\{D_\alpha(a, B), a \in A\}$$

$$D_\alpha(a, B) = Q_\alpha\{\rho(a, b), b \in B\}$$

He proposed values $\alpha = 1\%$ and $\beta = 10\%$. This prevents irrelevant points of $A$ or $B$ from altering the measure.

# Modified Hausdorff Distance

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 $\rho(a,b)$ be the Euclidean distance. They use

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

$$MHD(A,B) = \max\{h_{MHD}(A,B),\ h_{MHD}(B,A)\}, \tag{6}$$

and called this distance Modified Hausdorff Distance (MHD).They claim than it suites in best way the problem for object matching.

Similar approach (called Weighted Hausdorff Distance) is used by Lue, Tan, Huang and Fan (2001) for finding word image matching method in English and Chinese document images.

$$h_w(A,B) = \frac{1}{N_A} \sum_{a \in A} w(a) D(a,B), \text{ where } \sum_{a \in A} w(a) = N_A. \tag{7}$$

# Sum Hausdorff Distance

Two years ago in Ohrid we proposed to simplify MHD omitting the division by $N_A$, i.e.

$$h_{SHD}(A,B) = \sum_{a \in A} D(a,B), \tag{8}$$

and call the distance Sum Hausdorff Distance (SHD). This distance behaves pretty good for the purposes of word matching and the computer results shown in the Proceeding from Ohrid Conference were encouraging.

# Least Trimmed Square - HD

In 1999 D.-G. Sim, O.-K. Kwon, and R.-H. Park announced Least Trimmed Square - HD (LTS-HD) method for comparing noisy binary images.

$$h_\alpha(A, B) = \frac{1}{H} \sum_{i=1}^{H} D(a_i, B),$$

where $H = \alpha N_A$ and $D(a_1, B) \leq D(a_2, B) \leq \cdots \leq D(a_{N_A}, B)$. The parameter $\alpha$ could be choosen as $60 - 80\%$.

# Windowed Hausdorff Distance

(The latest idea)

In 2007 Baudrier, Nicolier, Millon and Ruan try to avoid the noise in the images and propose Windowed Hausdorff Distance (WHD).

If $W \in R^2$ and $Fr(W)$ is the boundary of $W$ then

$$WHD(A,B) = \max\{h_W(A,B),\ h_W(B,A)\}, \tag{9}$$

where there are three cases:

1. If $A \cap W \neq \emptyset$ and $B \cap W \neq \emptyset$

$$h_W(A,B) = \max_{a \in A \cap W} \{ \min_{b \in B \cap W} \rho(a,b), \min_{b \in Fr(W)} \rho(a,b) \};$$

2. If $A \cap W \neq \emptyset$ and $B \cap W = \emptyset$

$$h_W(A,B) = \max_{a \in A \cap W} \min_{b \in Fr(W)} \rho(a,b));$$

3. If $A \cap W = \emptyset$ then $h_W(A, B) = 0$.

The problem in the above definition is how the set $W$ to be chosen. The main difference with the classic HD is the term $\min_{b \in Fr(W)} \rho(a, b)$. To avoid the set $W$, the authors propose a parameter-free, adaptative, local Hausdorff distance. A window $W = B(x, r)$ is said to give a local measure at the point $x$ when the measure of the HD in the window $B(x, r)$ is maximum:

$$HD_{B(x,r)}(A, B) = r.$$

They defined Local Dissimilarity Map (LDMap):

$$LDMap(x) = \begin{cases} D(x, A) & \text{if } x \in A, \\ D(x, B) & \text{if } x \in B, \\ 0 & else \end{cases}$$

For comparing word images we need numbers, so that converting given LDMap to a number, we can use an appropriate norm. With $l_1$ norm the proposed method becomes equivalent to SHD.

# 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

$$\{W_i\}_{i=1,2,\ldots} \tag{10}$$

ordered according to a specific criteria. It can be some of the distances mentioned above.

For a given $n$ ($n = 1, 2, \ldots$), let $m(n) \leq n$ be the number of words among the first $n$ words of (10) that coincide with $W$. Then we define

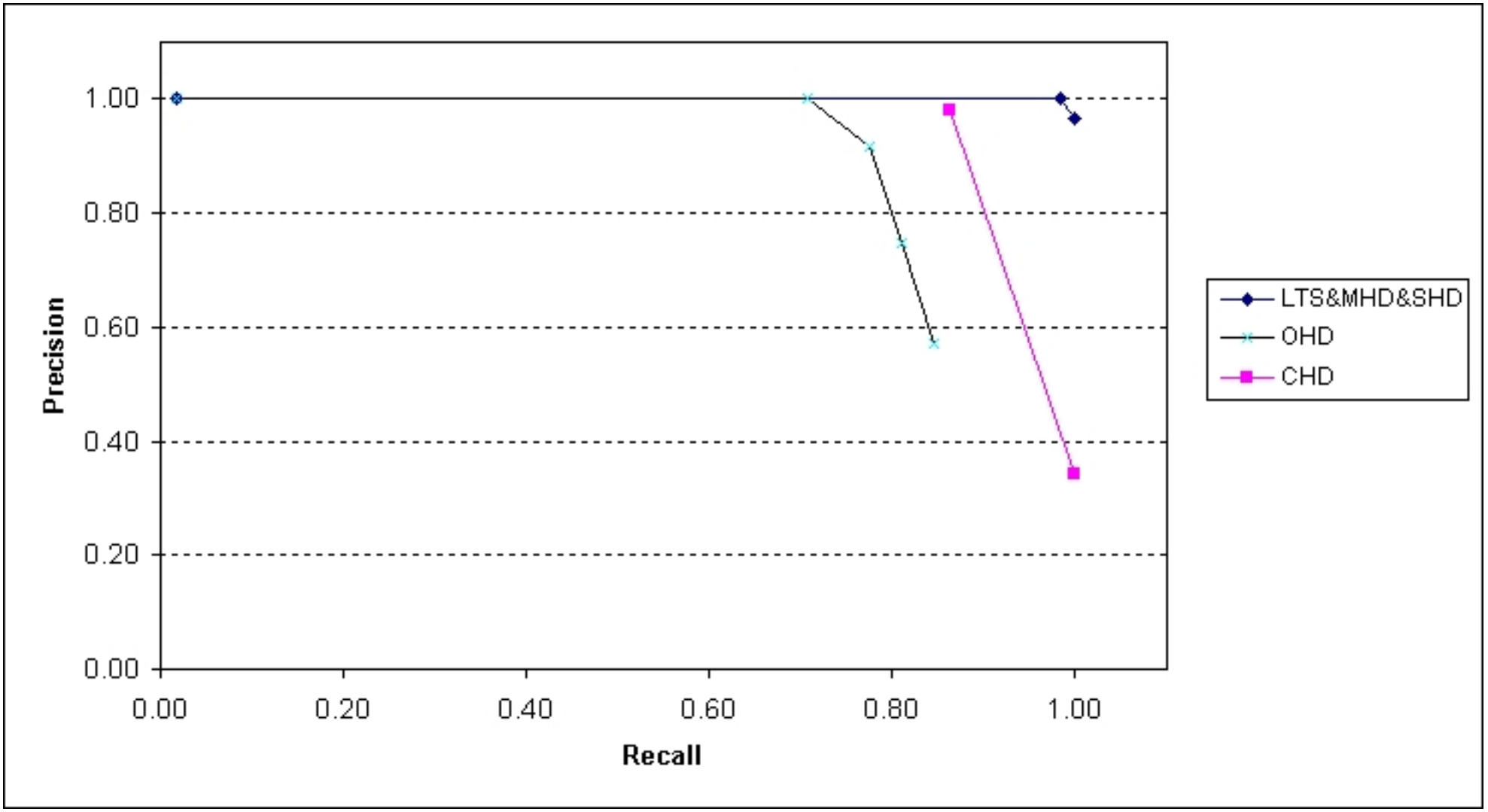
$$\textit{Recall}(n) = \frac{m(n)}{N} \quad \text{and} \quad \textit{Precision}(n) = \frac{m(n)}{n}. \tag{11}$$

# Experiments

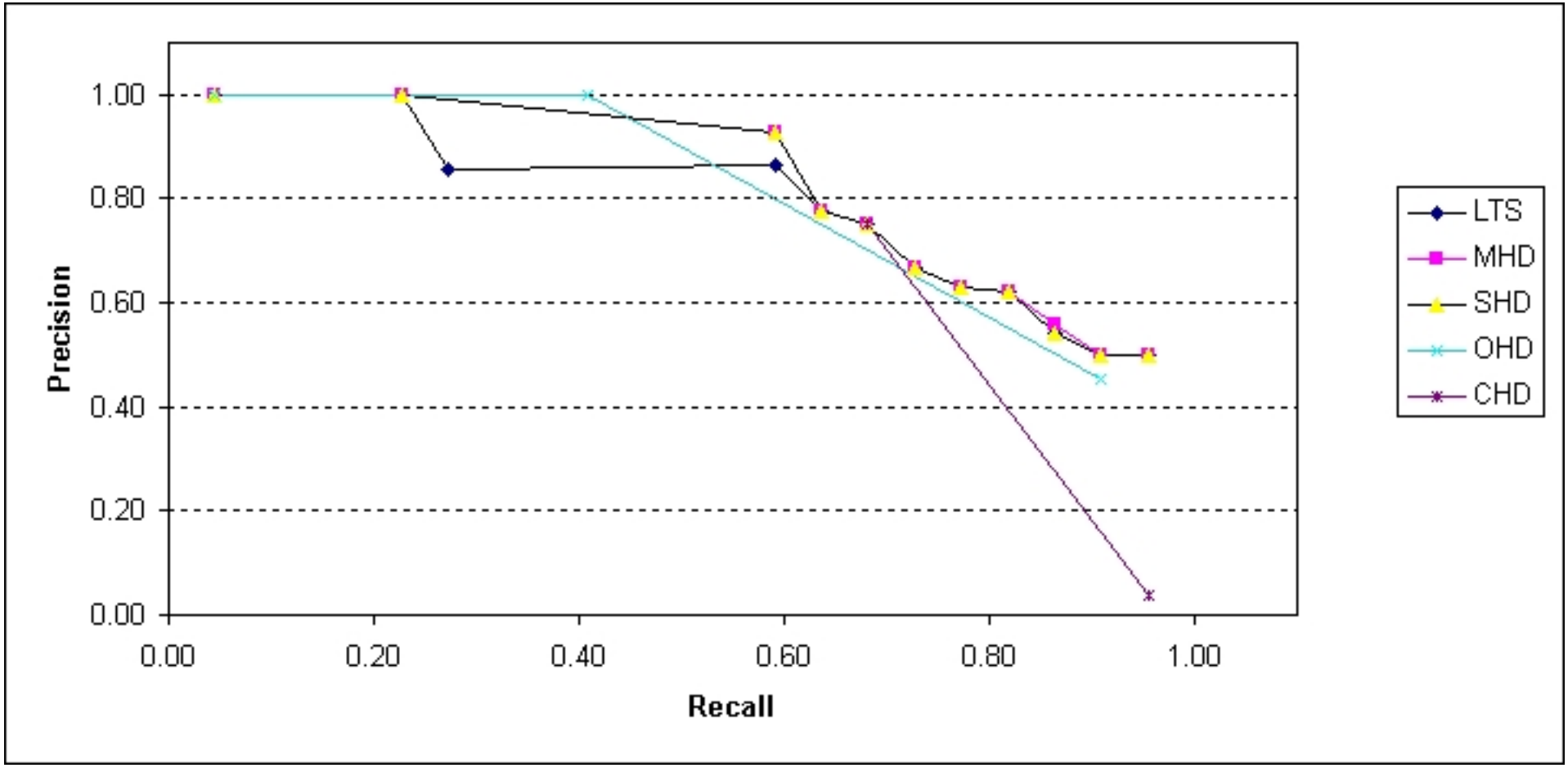
Using the distances defined above we carry out a series of computer word matching experiments. Real Bulgarian documents of typewritten low quality text of 54 pages are the material from which a specified word is located and extracted.

We present here 3 cases – for relatively large word (**Пазарджик**), for 5-letter word (**песни**) and for a short word (**така**).

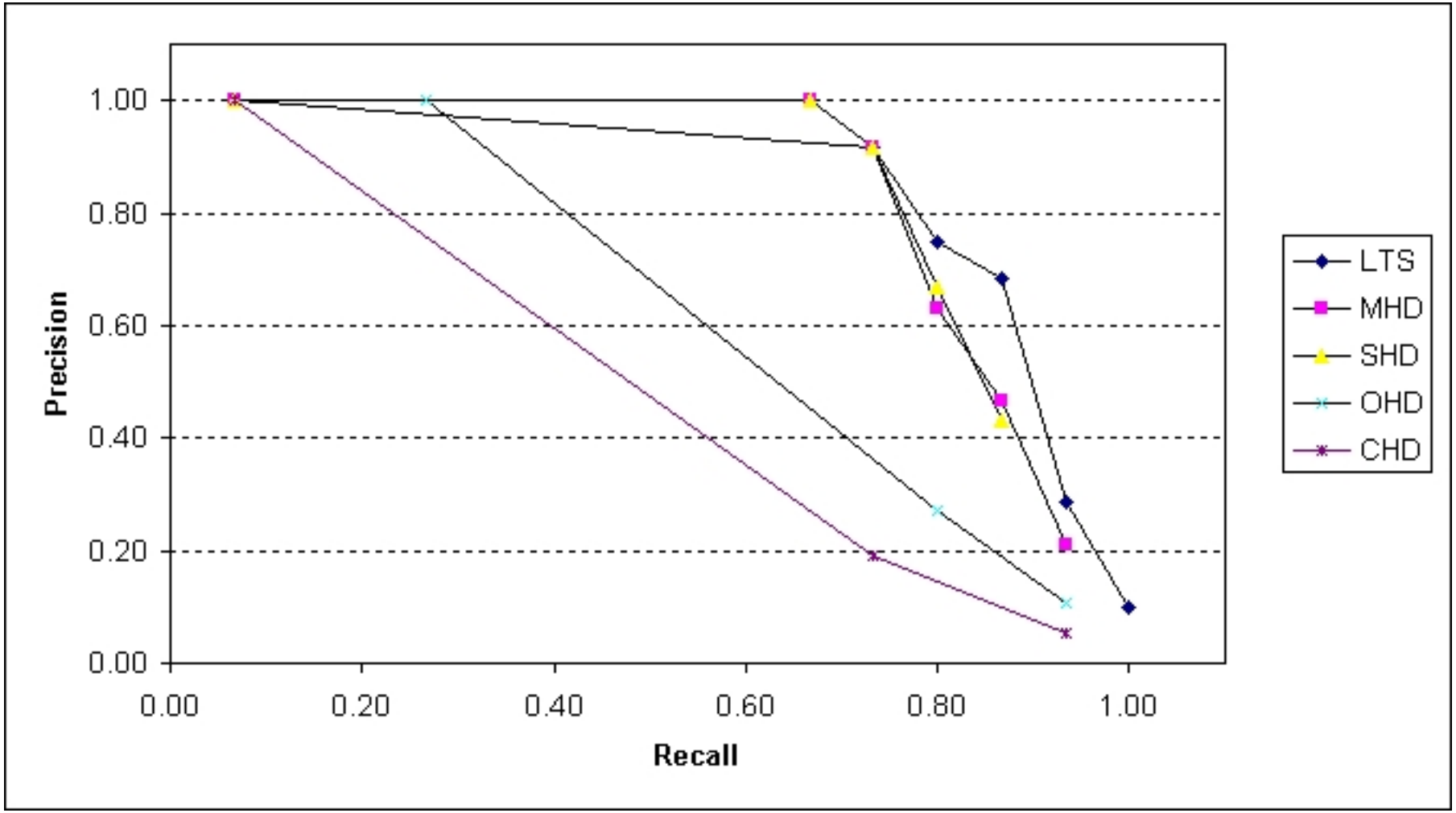
The word **Пазарджик** occurs 58 times.



The word **песни** occurs 22 times.



The word **така** occurs 15 times.



# Conclusion

We process low quality typewritten Bulgarian text for word matching using various distances. The results show that:

- The distances LTS, MHD and SHD produce almost the same results and therefore there is no need to complicate the definition of SHD.
- The measurement done by OHD and CHD could be considered as a "discontinuity". This explains the deterioration of the results produced these methods for values of $Recall(n) \approx 1$. For example, for the word **песни** with occurrence 31 times HD finds:

| OHD | $n$ | $m(n)$ |
|---|---|---|
| 3 | 9 | 9 |
| 4 | 35 | 11 |
| 5 | 136 | 2 |

| CHD | $n$ | $m(n)$ |
|---|---|---|
| 4 | 20 | 15 |
| 5 | 561 | 6 |

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

**Thank you for your attention.**

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

2. E. Baudrier, F. Nicolier, G. Millon, Su Ruan, *Binary-image comparison with local-dissimilarity quantification*, (Accepted in Pattern Recognition, July, 2007).

3. 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.

4. B. Gatos, I. Pratikakis and S. J. Perantonis, An Adaptive Binarization Technique for Low Quality Historical Documents, IARP Workshop on Document Analysis Syatems (DAS2004), Lectute Notes in Computer Science 3163, (2004) 102-113.

5. T. Konidaris, B. Gatos, K. Ntzios, I. Pratikakis, S. Theodoridis, S. J. Perantonis, *Keyword-guided word spotting in historical printed documents using synthetic data and user feedback*, International Journal of Document Analysis and Recognition, 9, (2007) 167Ц177.

6. 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.

7. 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.

8. José Paumard, *Robust comparison of binary images*, Pattern Recognition Letters 18 (1997) 1057-1063.

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