Преглед НЦД 12 (2008), 3–8

Andrey Andreev,
Nikolay Kirov
(Institute of Mathematics and Informatics,
Bulgarian Academy of Sciences)

SOME VARIANTS OF HAUSDORFF DISTANCE FOR WORD MATCHING [1]

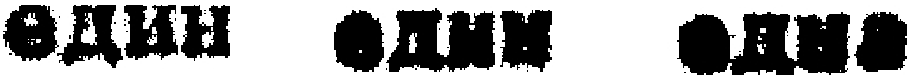
Abstract: Several recently proposed modifications of Hausdorff distance (*HD*) are examined with respect to word image matching for bad quality typewritten Bulgarian text. The main idea of these approaches presumes that omission of the extreme distances between the points of the compared images eliminates the noise (to some extent) and the algorithms become more robust. A few robust *HD* measures, namely, censored *HD, LTS-HD*, and a new binary image comparison method that uses a windowed Hausdorff distance, lie in the base of the computer experiments carried out using 54 pages of typewritten text.

Key words: text image, word matching, Hausdorff distance, letter searching

1. Introduction

The reasons to study the problem of word matching lie in:

- there exist a lot of binary images of scanned pages of low quality historical documents;
- it is difficult to OCR them because of low quality, old grammar and spelling, and old letters (not used nowadays);
- it is difficult even for a man to understand the letters and words in the text like these:



The application of the Hausdorff distance (*HD*) to the comparison of binary images presented in 1993 by Huttenlocher et al.[7] is used widely in different modified forms. In 2005 at Ohrid conference in article [1] some of the most popular varieties of *HD* were examined for their effectiveness in word matching. Typewritten text. of 49 pages of bad quality historical Bulgarian records was the material from which specifi words of different length were located and extracted.

The proposals for solving the problem are:

- to write a software tool which implements an efficient method for searching a word (a pattern word) in a set of pages (Gatos, Pratikakis, Perantonis, [6], 2004);
- the base of such a tool is a method for word comparison, and after that ordering words with respect to their distances according to a chosen pattern word;
- the pattern word can be a synthetic keyword (Konidaris, Gatos and al. [9], 2007) or can be a real word chosen from a text (feedback approach).

[1] This research has been partially supported by a Marie Curie Fellowship of the European Community program "Knowledge Transfer for Digitalization of Cultural and Scientific Heritage in Bulgaria" under contract number MTKD-CT-2004-509754.

This paper can be considered as a continuation of [1] in the field of application of *HD* to word matching. It follows the ideas of many authors to find out robust modification of *HD*.

Particularly for bad quality text the robustness of *HD* allows decreasing the influence of the noise and leads to better results. It can be summarized that all methods considered further exploit the idea of ignoring some amount of the extreme distances between the points of the images supposing that such points are noise.

Let A and B denote bounded sets on the Euclidean plane R^2 and let a and b be points on R^2 with coordinates $a = (a_1, a_2)$, $b = (b_1, b_2)$. Then *HD* between A and B is

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

where

$$h(A,B) = \max_{a \in A} \min_{b \in B} \rho(a,b), \quad \rho(a,b) = \max\{|\, a_1 - b_1 \,|, |\, a_2 - b_2 \,|\}.$$

Instead of the above definition of $\rho(a,b)$ an arbitrary point distance in the plane R^2 can be used.

Five different distances were examined numerically in [1] for the purpose of word matching. Among them is so called “Modified Hausdorff Distance” (*MHD*). Dubuisson and Jain [5] proved numerically that the “distance” *MHD* produces the best results among the 24 distances of Hausdorff type. In [1] *MHD* was slightly simplified and *SHD* distance

$$SHD(A,B) = \max\{h(A,B),\, h(B,A)\},\ h(A,B) = \sum_{a \in A} \min\{\rho(a,b) : b \in B\}$$

was proposed. Computer experiments show that for our goals *MHD* and *SHD* behave equivalently (with tiny advantage for *SHD*) and that they surpass the other methods.

2. Robust Hausdorff distances used in word matching

2.1. Censored Hausdorff distance $CHD_{\alpha,\beta}(A,B)$. The idea of Azencott et al. [2] and Paumard [10] is: for a point a from A the p closest neighbors of a in B must not be considered, where $p=\alpha N_B$, where N_B denotes the number of points of the set B and $0 \le \alpha \le 1$. Namely, let the set $X = \{x_i\}$ consist of real numbers x_i for which

$$x_1 \le x_2 \le \ldots \le x_{N_X}$$

and let

$$X_\alpha = \alpha N_X \ , \ Q_\alpha\{X\} = x_{X_\alpha} .$$

For $0 \le \alpha, \beta \le 1$,

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

where

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

The proposed values for α and β that prevents irrelevant points from A and B to alter the measure are $\alpha = 0.01$, $\beta = 0.1$.

2.2. Least Trimmed Square distance *LTS-HD*(*A,B*). In 1999 new robust variation of *HD* was proposed in [4] for the purposes of image matching

$$LTS\text{–}HD_{\alpha}(A,B) = \max\{h_{\alpha}(A,B)\,,\, h_{\alpha}(B,A)\}\,,$$

where

$$h_{\alpha}(A,B) = \frac{1}{\alpha N_A} \sum_{i=1}^{\alpha N_A} \min\{\rho(a_i,b) : b \in B\}$$

and

$$D(a_1,B) \le D(a_2,B) \le \ldots \le D(a_{N_A},B)\,, \quad D(a,B) = \min\{\rho(a,b) : b \in B\}\,.$$

The suggested value for the parameter α is 0.8.

2.3. Windowed Hausdorff distance *WHD*$_W$(*A,B*). In January 2007 in a preprint (later published in [3]) new approach was given for avoiding the noise in the images. The window *W* is a set in R^2 and *Fr*(*W*) is the boundary of *W* then

$$WHD_W(A,B) = \max\{h_W(A,B)\,,\, h_W(B,A)\},$$

where:

- if $A \cap W \neq \emptyset$ and $B \cap W \neq \emptyset$ then

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

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

- if $A \cap W \neq \emptyset$ and $B \cap W = \emptyset$ then

- if $A \cap W = \emptyset$ then $h_W(A,B) = 0$.

The problem in the above definition is how the window *W* to be chosen and the main difference with the classic definition of *HD* is the term

$$\min_{b \in Fr(W)} \rho(a,b)\,.$$

The authors in [3] introduced the notion of local dissimilarity claiming that a ball *B*(*x,r*) gives a local *HD* between the sets *A* and *B* in the window *B*(*x,r*) when

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

They also defined a Local Dissimilarity Map (*LDMap*)

$$LDMap(x) = \begin{cases} \min\{\rho(x,b)\,,\, b \in B\} & \text{if } \; x \in A\,,\, x \notin B, \\ \min\{\rho(x,a)\,,\, a \in A\} & \text{if } \; x \in B\,,\, x \notin A, \\ 0 \quad \text{otherwise}\,. \end{cases}$$

For comparing word images we need to convert a given *LDMap* into a number using an appropriate norm. With the discrete L_1 norm of the function *LDMap*(*x*) we obtain *SHD*.

3. Experimental results

3.1. Measuring the effectiveness of the methods. The effectiveness of the distances is given usually by the standard estimations Recall and Precision – Junker, Hoch and Dengel [8]. Let us look for a word *W* in a collection of binary text images in which *W* occurs *N* times. According to a specific criteria let a given method produce a sequence of words $\{W_i\}_{i=1,2,...}$. Such criteria can be some of the distances mentioned above. For a fixed $n = 1, 2,...$, let $m(n) \leq n$ be the number of words among the first *n* words in the sequence $\{W_i\}_{i=1,2,...,}$that coincide with *W*. Then we define

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

3.2. Numerical results We have tested numerically the effectiveness of the distances defined above in comparison with the distances *MHD* and *SHD*, which are considered nowadays among the best. Bulgarian documents of typewritten text of 54 pages of bad quality are the material from which three words of different length are located and extracted.

The results for relatively long word **Пазарджик** which occurs 58 times in the text are shown on Figure 1. Figures 2 and 3 reprisent the results for the words **песни** and **така**. They occur respectively 22 and 15 times in the text. In these figures the distances *HD* and *LTS-HD* are denoted by *OHD* and *LTS* respectively.

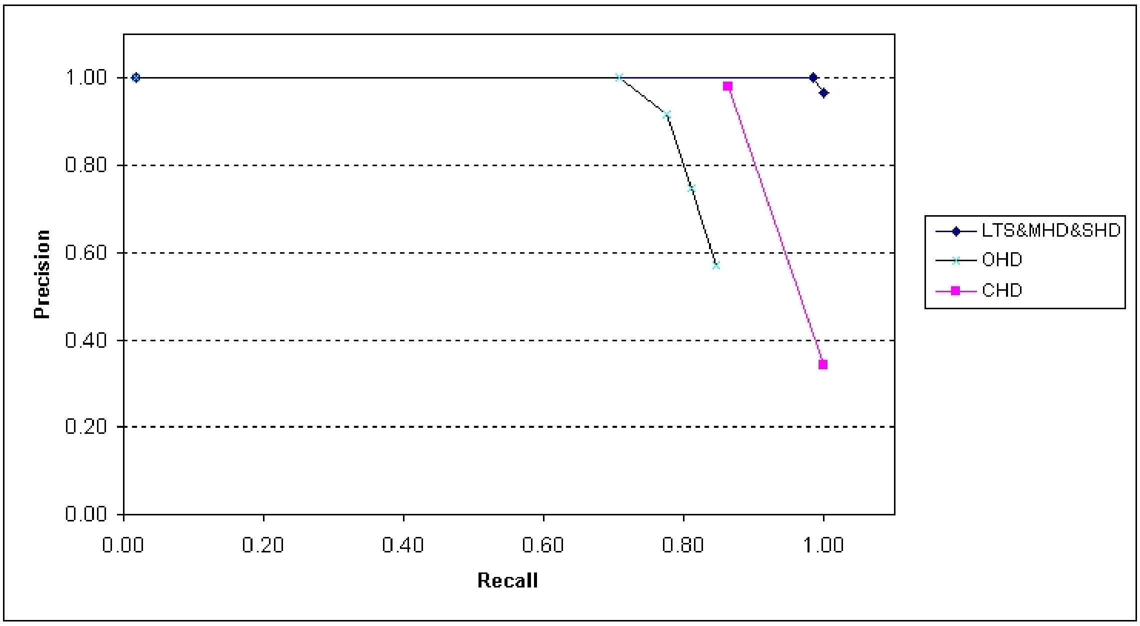


Fig. 1. Results for **Пазарджик**

4. Conclusions

- The distances *LTS-HD*, *MHD* and *SHD* produce almost the same results. The reason is that these distances accumulate the distances of all points (in *LTS-HD* almost all points) avoiding in this way the discrete narrow scale of measuring of *HD* and *CHD*.

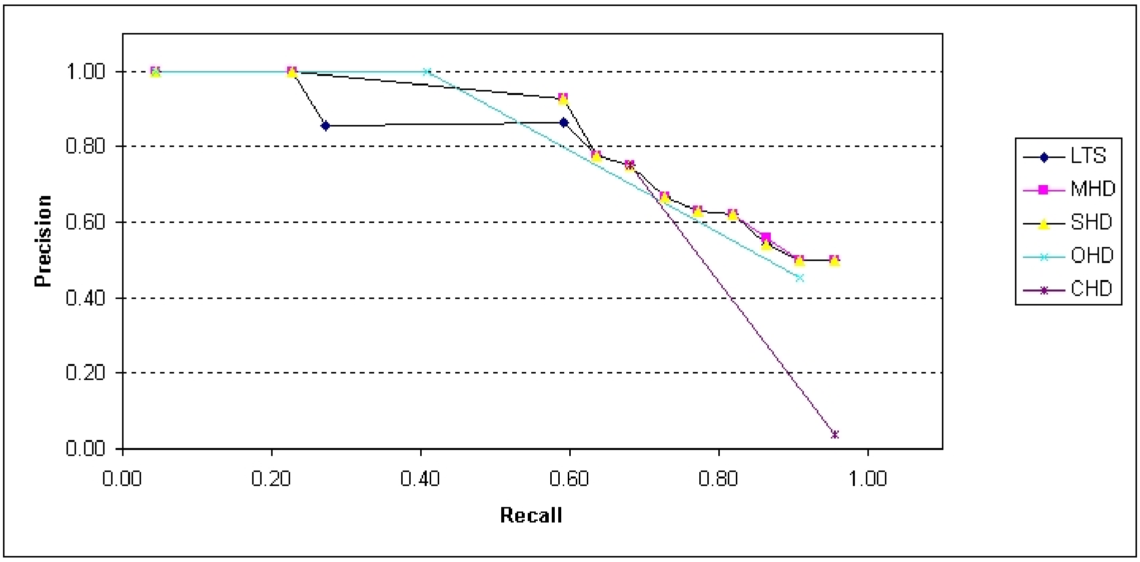


Fig. 2. Results for **песни**

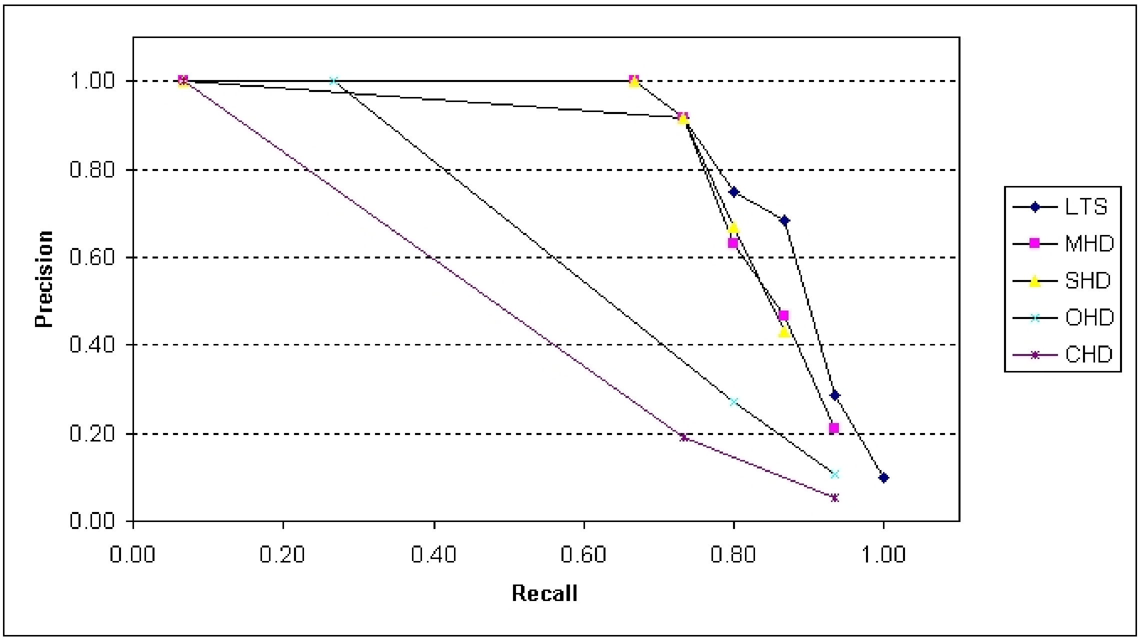


Fig. 3. Results for **така**

- *HD* and *CHD* could be considered as a "discontinuity". They operate relatively small amount of integer numbers for representation of the distances between the sets. In the case of word matching these integers are equal to 3, 4 or 5. This explains the deterioration of the results produced by these methods for values of Recall(n) closed to 1. For example, for the word **песни** with occurrence 22 times *HD* and *CHD* provide the following results:

HD	*n*	*M*(*n*)		*CHD*	*n*	*m*(*n*)
3	9	9				
4	35	11		4	20	15
5	136	2		5	561	6

In this sense the methods *MHD*, *SHD* and *LTS-HD* practically use continuous scale for ordering the spotted words ensuring much better results.

References

[1] A. Andreev, N. Kirov, *Word image matching in Bulgarian historical documents*, [SEEDI Communications 1], Review of the National Center for Digitization, 8 (2006), 29–35.

[2] R. Azencott, F. Durbin, J. Paumard, *Multiscale identification of building in compressed large aerial scenes*, in Proc. 13th Int. Conf. Pattern Recognition, Vienna, Austria, Aug., 1996, vol. 2, 974–978.

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

[4] Dong-Gyu Sim, Oh-Kyu Kwon, Rae-Hong Park, *Object Matching Algorithms Using Robust Hausdorff Distance Measures*, IEEE Transactions on Image Processing, vol. 8, No. 3, March 1999, 425–429.

[5] 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.

[6] B. Gatos, I. Pratikakis and S. J. Perantonis, *An Adaptive Binarization for Low Quality Historical Documents.* IARP Workshop on Document Analysis System (DAS2004), Lecture Notes in Computer Science, 3163, (2004), 102–113.

[7] D. P. Huttenlocher, G. A. Klanderman, J.R. William, *Comparing images using the Hausdorff distance*. IEEE Trans. Pattern Anal. Machine Intell. 15(9), (1993)

[8] 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.

[9] T. Konidaris, B. Gatos, K. Nitzios, I. Pratikakis, S. Theodoridis, S. J. Perantonis, *Keyword-guided word spotting in historical printed documents using synthetic data and user feedback.* International Journal on Document Analysis and Recognition, 9, (2007), 167–177.

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

aandreev@math.bas.bg
nkirov@math.bas.bg