Text Search in Document Images Based on Hausdorff Distance Measures

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

Abstract: *The Hausdorff type distances between the sets of points on the plane are the commonly used similarity measures for binary images. In this work we present several such measures in a unified manner and introduce a new, naturally arisen variant of Hausdorff distance. The matching performance of all similarity measures is compared by computer experiments, using real word images from a scanned book.*

Key words: *Binary Text Images, Hausdorff Distance, Similarity Measures, Word Searching*

Introduction

Libraries contain huge amounts of historical documents which cannot be made available online because they do not have a searchable index. The wordspotting idea has been proposed as a solution for creating indexes for such documents by matching word images. Optical character recognition is the usual way of conducting text retrieval from scanned document images. Moreover recognizing full text in images is a wasteful task for information retrieval. The motivation of our work is to choose effective search in scanned documents by simply considering the image similarities. One of the most widespread ideas is to use Hausdorff type measures for word image similarity.

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

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

where $h(A,B)$ and $h(B,A)$ are co-called directed distances between the sets. For original Hausdorff metrics

$$h(A,B) = \max_{a \in A} d(a,B), \text{ where } \quad d(a,B) = \min_{b \in B} \rho(a,b),$$

i.e. $d(a,B)$ is the distance from a point a to the set B, and $\rho(a,b)$ is a point distance.

Euclidean distance: $\rho(a,b) = \sqrt{(a_x - b_x)^2 + (a_y - b_y)^2}$.

Manhattan distance: $\rho(a,b) = |a_x - b_x| + |a_y - b_y|$.

Infinity norm distance: $\rho(a,b) = \max\{|a_x - b_x|, |a_y - b_y|\}$.[2pt] 0-1 distance:

$$\rho(a,b) = \begin{cases} 0 & \text{if } \ a \equiv b \\ 1 & \text{otherwise} \end{cases} \tag{2}$$

Huttenlocher *at al.* proposed the Partial Hausdorff Distance (PHD) for comparing images containing a lot of degradations or occlusions. For directed distance they take the K-th ranked point of A instead of the largest one

$$h_K(A, B) = K^{th}_{a \in A} d(a, B), \tag{3}$$

where $K^{th}_{a \in A}$ denotes the K-th ranked value in the set of distances $\{d(a, B) : a \in A\}$, 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 values to this distance,

$$d(a_1, B) \geq d(a_2, B) \geq \cdots \geq d(a_K, B) \geq \cdots \geq d(a_{N_A}, B). \tag{4}$$

This HD measure requires one parameter, often represented by $f = K/N_A$ ($0 \leq f \leq 1$). Sim *at al.* claim that a value in the interval $[0.6, 0.8]$ gives good matching results. Note that this measure is not a metric because $h_K(A, A) > 0$!

The idea of José Paumard is that we do not take into account the L closest neighbors of $a \in A$ in B. So we can 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 $L^{th}_{b \in B}$ denotes the L-the ranked value in the set of distances $\{\rho(a, b) : b \in B\}$ for a given point a of A. Now the directional Censored Hausdorff Distance (CHD) can be defined as

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

Let us set two parameters $\alpha = K/N_A$ and $\beta = L/N_B$ which are relative values with respect to the number of points in the sets A and B. Then the recommended values in for these parameters are $\alpha = 0.1$ and $\beta = 0.01$.

For all three described measures (HD, PHD and CHD), the directed distance can be considered as a choice a representative pair of points (a_0, b_0), $a_0 \in A$ and $b_0 \in B$ such that the point distance between them $\rho(a_0, b_0)$ is equal to the corresponding directed distance between the sets A and B.

Another approach for measuring similarity between two finite sets in the plane is to calculate a sum of point distances.

Dubuisson and Jain examined a number of distance measures of Hausdorff type for determination to what extend two point sets on the plane A and B differ. They introduced so-called Modified Hausdorff Distance (MHD) with the following distance measure

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

They claim than it suites in best way the problem for object matching.
A bit better results were obtained in our examples omitting the coefficient $1/N_A$ in front of the sum. We called this modification Sum Hausdorff Distance (SHD)

$$h_{\mathsf{SHD}}(A,B) = \sum_{a\in A} d(a,B) = \sum_{a\in A}\min_{b\in B}\rho(a,b). \tag{7}$$

In 1999 D.-G. Sim *at al.* described two variants 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 distances.
The directed distance for M-HD is defined by

$$h_{\mathrm{M}}(A,B) = \frac{1}{N_A} \sum_{a \in A} f(d(a,B)), \tag{8}$$

where the function f is convex and symmetric and has a unique minimum value at zero. One possible function is

$$f(x) = \begin{cases} |x| & \text{if } |x| \leq \tau \\ \tau & \text{if } |x| > \tau \end{cases}$$

This means that we sum the distances $d(a,B)$ which are less than the constant τ and add τ when the distance is greater than τ. The recommended interval of τ is $[3,5]$. Note that MHD with 0-1 point distance is M-HD for $\tau = 1$.

The second measure is called Least Trimmed Square HD (LTS-HD). The directed distance is

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

where $K \leq N_A$ and $a_1, a_2, \ldots, a_{N_A}$ are points of A for which (4) is valid. Parametrization of the method can be done by a parameter $\alpha = K/N_A$. For comparing noisy binary images the suggested value for this parameter is 0.2.

Following the definition of CHD, we introduce its analogical method based on the sum of point distances. The directed distance is

$$h_{\mathsf{NEW}}(A,B) = \frac{1}{N_A - K} \sum_{i=K}^{N_A} d_L(a_i, B) = \frac{1}{N_A - K} \sum_{i=K}^{N_A} L^{th}_{b \in B} \rho(a,b). \tag{10}$$

We can set again the parameters $\alpha = K/N_A$ and $\beta = L/N_B$ which are relative values with respect to the number of points in the sets A and B.

A new approach to similarity measures

We can consider a linear order of points of A and give a sequence representation: $A = \{a_1, a_2, \ldots, a_{N_A}\}$. For every $a_k \in A$ $(k = 1, 2, 3, \ldots, N_A)$ we can calculate the distances (with respect to a metric ρ in R^2) from a_k to all points in B, i.e.

$$d_k^1 = \min_{b \in B} \rho(a_k, b) = \rho(a_k, b_k^1), \quad d_k^2 = \min_{b \in B \setminus \{b_k^1\}} \rho(a_k, b) = \rho(a_k, b_k^2), \ldots,$$

$$d_k^l = \min\{\rho(a_k, b) : b \in B \setminus \{b_k^1, b_k^2, \ldots b_k^{l-1}\}\} = \rho(a_k, b_k^l), \ldots,$$

obtaining in such a way a nondecreasing sequence of numbers

$$d_k^1 \leq d_k^2 \leq \cdots \leq d_k^l \leq \cdots \leq d_k^{N_B}.$$

Carrying out these calculations for every point in A, we define a distance matrix D

$$D = \begin{pmatrix} d_1^1 & d_1^2 & d_1^3 & \dots & d_1^l & \dots & d_1^{N_B} \\ d_2^1 & d_2^2 & d_2^3 & \dots & d_2^l & \dots & d_2^{N_B} \\ d_3^1 & d_3^2 & d_3^3 & \dots & d_3^l & \dots & d_3^{N_B} \\ \dots & \dots & \dots & & \dots & & \dots \\ d_k^1 & d_k^2 & d_k^3 & \dots & d_k^l & \dots & d_k^{N_B} \\ \dots & \dots & \dots & & \dots & & \dots \\ d_{N_A}^1 & d_{N_A}^2 & d_{N_A}^3 & \dots & d_{N_A}^l & \dots & d_{N_A}^{N_B} \end{pmatrix}$$

following arbitrary order of points in A. Later we will choose ordering of rows, corresponding to an order in a column. For definition of MHD and M-HD we do not need any order

$$h_{\mathsf{MHD}}(A,B) = \frac{1}{N_A}\sum_{i=1}^{N_A} d_i^1, \quad \text{and} \quad h_{\mathsf{M}}(A,B) = \frac{1}{N_A}\sum_{i=1}^{N_A} \min\{d_i^1, \tau\}.$$

For finding the Hausdorff distance in the distance matrix D, we consider the following order (obtained by swapping the rows) with respect to the first column of D

$$h(A,B) = d_1^1 \geq d_2^1 \geq \cdots \geq d_k^1 \geq \cdots \geq d_{N_A}^1.$$

The directed distance for PHD is $h_K(A,B) = d_K^1$. We can calculate LTS-HD summing the part of the first column elements

$$h_{\mathsf{LST}}(A,B) = \frac{1}{N_A - K} \sum_{i=K}^{N_A} d_i^1.$$

We can find CHD directed distance as an element of matrix D swapping the matrix rows in such way that the L-th column is sorted, i.e.

$$d_1^L \geq d_2^L \geq \cdots \geq d_k^L \geq \cdots \geq d_{N_A}^L.$$

Then $h_{K,L}(A,B) = d_K^L$. The directed NEW distance is

$$h_{\mathsf{NEW}} = \frac{1}{N_A - K} \sum_{i=K}^{N_A} d_i^L.$$

Experiments

We carried out our experiments using an old book (1884) – Bulgarian Chrestomathy, created by famous Bulgarian writers Ivan Vasov and Konstantin Velichkov. The quality of scanned images are quite bad because this was one of the first books, processing in the digitization center and operators' qualification was not on appropriate level. Many pages have slopes in rows, there are significant variations in gray levels, etc.

There is no text version till now of this book, which may be produced using appropriate OCR software. The first reason is the quality of images. The second reason is the absence of OCR software because the text contains old and abandoned Bulgarian letters. Also spelling and grammar are quite different in modern Bulgarian language.

поетъ, сатирикъ и публицистъ. Първо-то нѣщто, което е издалъ е книжка стихотворения „Басненникъ“ и по-послѣ „Смѣсна Китка“ (Букурештъ 1852 г.), съ които той доби първа-та си извѣстность у насъ, като български писатель. Отъ 1857 год. се почева негова та многополезна дѣятелность въ борба-та ни съ Гръци-тѣ за черковна независимость. Той дохожда въ Цариградъ и издава свои-тѣ „Смѣшни Календари“ сатирически книги, въ които бичува съ единъ искусенъ и ядовитъ сарказмъ пороци-тѣ и недостатки-тѣ на тогавашно-то българско обштество, и гръцко-то високо духовенство (1857—1863). На 1863 год. той прѣдприе издаване-то на сатирический вѣстникъ „Гайда,“ който не трая много врѣме. Доста хубави статии все въ полемическо-сатирический духъ, напечата той тамъ. Слѣдъ двѣ години Славейковъ прѣдприе издание-то на политический вѣстникъ „Македония“ (1867—1870). Тамъ при разискване-то на разни въпроси отъ обштественъ и черковенъ интересъ Славейковъ се стараеше да разбуди народно-то чувство у Македонски-тѣ Българе, които душеше нетърпимо-то влияние на гръкоман, ство-то и фанариотство-то. Най-послѣ подиръ нѣколко врѣменни спирания и конфискации на вѣстникъ-тъ, правителство-то съвсѣмъ го уништожи и запрѣти на Славейкова да издава вече какъвъ-да-е вѣстникъ, а и него самаго тури въ тъмница, по обвинение, че въ послѣдни-тѣ броеве на „Македония“ явно проповѣдвалъ революционни идеи между Българе-тѣ.

We used 200 pages from about 1000 book pages scanned at a resolution of 200 DPI. The images are about 2300×3600 pixels (8.28 MPixels), 14.8 x 23.3 cm, grayscale 256 (8 BitsPerPixel). We use preprocessing to convert the images to 1 bit per pixel, black and white, by the help of Image Magic software with 60% threshold value.

The goal of our experiments is to compare practically the efficiency of described methods counting the number of correctly retrieved words in a sequence of words, sorted by their similarity measures with respect to the corresponding HD. For all experiments the same segmentation is used. We choose a pattern word and then measure similarities between it and the words with approximately same width.

Tables contains numbers of correct words in an ordered sequence with the corresponding distance D. m and n in the ratio m/n denote:
– m, the number of correct words with distance D;
– n, the number of all words with distance D.

For word **всички**

$D =$ / Method	4	5	6	7	8
HD	16/16	44/44	115/120	168/217	177/500
PHD+3	**77/77**	206/254	209/500	–	–
CHD	19/19	**213/252**	**214/500**	–	–

For word **Русия**

$D =$ / Method	4	5	6	7
HD+1	2/2	3/3	5/5	5/6
PHD+3	3/3	**11/15**	–	–
CHD		**8/8**	**13/24**	–

We count the number of correctly retrieved words among first $100, 200, \ldots, 500$ words with approximately same width. m is the number of correctly retrieved words among first n words in the ordered sequence in the notation m/n.

For word **всички**

$n =$ Method	100	200	300	400	500
HD01	97	158	186	195	206
MHD	**100**	169	199	207	212
SHD	**100**	177	205	213	220
M-HD	**100**	173	202	214	218
LTS-HD	**100**	**185**	**215**	**221**	**224**
NEW	97	164	198	213	**224**

For word **Русия**

Method			
HD01	4/4	9/18	10/23
MHD	10/10	**14/23**	**15/49**
SHD	**11/11**	14/24	–
M-HD	7/7	12/14	–
LTS-HD	10/10	**14/23**	–
NEW	7/7	12/15	14/26

There are two relative words (derivatives) of the pattern word **всички**, namely **всичка** and **всичко**. We count as correct words all three of them. This is very useful in practice and show another advantage of methods under discussion and our approach in search. Also, there are 5 similar words of the word **Русия**: **Руски**, **Руска**, **Руско**, **руски** and **руска**.

The best results are in bold in all tables.

Discussion and Conclusion

In this article we do not discuss the quality of image preprocessing particularly the important step of segmentation. Also we have no data of number of searching words in the text, because this is tedious work which cannot be done by computer. It follows than we cannot produce the standard recall/precision retrieval estimation. In addition, we cannot catch the words which are incorrect segmented as well as these which are break at the end of a line and remaining part is placed on the next line. Nevertheless we think that our comparison of similarity methods is significant for their implementations in software searching systems. In spite of low efficiency of these Hausdorff type methods (the searching takes a lot of time) we believe that the modern, high level personal computers could be able to solve the problem in reasonable time.

The main conclusions that we derive from are:

1. “Sum-distances” outmatch “point-distances”.

2. There are no significant differences between the methods that we call “sum-distances” ones.

Thank you for your attention.