Word Image Matching Based on Hausdorff Distances

Andrey Andreev\textsuperscript{1} and Nikolay Kirov\textsuperscript{1,2}

\textsuperscript{1}Institute of Mathematics and Informatics, BAS
\textsuperscript{2}New Bulgarian University,
Sofia, Bulgaria

aandreev@math.bas.bg, nkirov@nbu.bg
Hausdorff distance (HD) and its modifications provides one of the best approaches for matching of binary images. This paper proposes a formalism generalizing almost all of these HD based methods. Numerical experiments for searching words in binary text images are carried out with old Bulgarian typewritten text, printed Bulgarian Chrestomathy from 1884 and Slavonic manuscript from 1574.

The main goals of this paper are:

- to propose a new method for estimating the similarity between two binary images in order to generalize and to unify the existing image matching methods based on Hausdorff distance;

- to check numerically the efficiency of generalized HD method when it is applied for word matching in typewritten, printed and handwritten historical documents.
Hausdorff distances for set similarities

The Hausdorff distance (HD) between two closed and bounded subsets $A$ and $B$ of a given metric space $M$ is defined by

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

where $h(A, B)$ is so-called directed distance from $A$ to $B$. For classical Hausdorff distance

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

de is the distance from a point $a$ to the set $B$, and $\rho(a, b)$ is a point distance in the metric space $M$.

D. P. Huttenlocher et al. proposed Partial Hausdorff Distance (PHD) for comparing images containing a lot of degradation or occlusions. Let $K^{th}_{a \in A}$ denote the $K$-th ranked value in the set of distances $\{d(a, B) : a \in A\} = \{d(a_i, B), i = 1, \ldots, N_A\}$, 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 distance values:

$$d(a_1, B) \geq \cdots \geq d(a_K, B) \geq \cdots \geq d(a_{N_A}, B). \quad (3)$$

This definition of $K^{th}_{a \in A}$ differs from the original one, where the rating order in (3) is in the opposite direction. The directed distance for PHD is

$$h_K(A, B) = K^{th}_{a \in A}d(a, B) = d(a_K, B). \quad (4)$$
The idea of J. Paumard is that we do not take into account the $L$ closest neighbours of $a \in A$ in $B$. So we define the distance from a point $a \in A$ to the set $B$ as follows

$$d_L(a, B) = L_{b \in B}^{th} \rho(a, b),$$

where $L_{b \in B}^{th} \rho(a, b) = \rho(a, b_L)$ denotes the $L$-th ranked value in the set of distances $\{\rho(a, b) : b \in B\} = \{\rho(a, b_i), i = 1, \ldots, N_B\}$, i.e.

$$\rho(a, b_1) \leq \cdots \leq \rho(a, b_L) \leq \cdots \leq \rho(a, b_{N_B}).$$

Now the directed Censored Hausdorff Distance (CHD) is defined by

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

M.-P. Dubuisson and A. Jain examined 24 distance measures of Hausdorff type to determine to what extent two finite sets $A$ and $B$ on the plane differ. Based on numerical behavior of these distances on synthetic images containing various levels of noise they introduced Modified Hausdorff Distance (MHD) with directed distance

$$h_{MHD}(A, B) = \frac{1}{N_A} \sum_{a \in A} d(a, B) = \frac{1}{N_A} \sum_{a \in A} \min_{b \in B} \rho(a, b). \quad (6)$$
In 1999 D.-G. Sim et al. described two modifications 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-HD. The directed M-HD is defined by

\[ h_M(A, B) = \frac{1}{N_A} \sum_{a \in A} f_{\tau}(d(a, B)), \]

where the function \( f_{\tau} : R_+ \to R_+ \) is increasing and has an unique minimum value at zero. They introduce one simple function with these properties

\[ f_{\tau}(x) = \min\{x, \tau\}, \]

for a given \( \tau > 0 \).

The directed distance of LTS-HD is defined by

\[ h_{\text{LTS}}(A, B) = \frac{1}{N_A - K + 1} \sum_{i=K}^{N_A} d(a_i, B), \]

where \( 1 \leq K \leq N_A \) and \( a_1, a_2, \ldots, a_{N_A} \) are the points of \( A \) for which (3) is valid.
A new approach to HD similarity measures

Let us suppose there is a linear order of the points of the set $A = \{a_1, a_2, \ldots, a_{N_A}\}$. For every $a_k \in A$ we calculate the distances from $a_k$ to all points in $B$, as follows:

$$
\begin{align*}
    d_{k1} &= \min_{b \in B} \rho(a_k, b) = \rho(a_k, b_{k1}), \\
    d_{k2} &= \min_{b \in B \setminus \{b_{k1}\}} \rho(a_k, b) = \rho(a_k, b_{k2}), \\
    \vdots, \\
    d_{kl} &= \min_{b \in B \setminus \{b_{k1}, \ldots, b_{kl-1}\}} \rho(a_k, b) = \rho(a_k, b_{kl}), \\
    \vdots
\end{align*}
$$

In such a way we obtain a nondecreasing sequence of non-negative numbers

$$
d_{k1} \leq d_{k2} \leq \cdots \leq d_{kl} \leq \cdots \leq d_{kN_B}.
$$

Let the matrix $D$ be defined by

$$
D = \begin{pmatrix}
    d_{11} & d_{12} & \cdots & d_{1l} & \cdots & d_{1N_B} \\
    \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\
    d_{k1} & d_{k2} & \cdots & d_{kl} & \cdots & d_{kN_B} \\
    \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\
    d_{N_A1} & d_{N_A2} & \cdots & d_{N_Al} & \cdots & d_{N_AN_B}
\end{pmatrix}.
$$

For a given $1 \leq l \leq N_B$, we define a new matrix $D_l$:

$$
D_l = \left(d_{ij}^{(l)}\right), \ i = 1, \ldots, N_A, \ j = 1, \ldots, N_B
$$
interchanging the rows of the matrix $D$ so that the elements of $l$-th column are sorted, i.e. satisfying the following inequalities:

$$d_{1l}^l \geq d_{2l}^l \geq \cdots \geq d_{kl}^l \geq \cdots \geq d_{N_A l}^l.$$ 

Let $1 \leq k \leq N_A$ and $1 \leq l \leq N_B$ be integer numbers. We define two Generalized Hausdorff Distances (GHD) using the following directed distances:

$$h_{k,l}^{(p)}(A, B) = d_{kl}^l$$

(11) and

$$h_{k,l}^{(s)}(A, B) = \frac{1}{N_A - k + 1} \sum_{i=k}^{N_A} d_{il}^l.$$ 

(12)

We denote (11) by p-GHD and (12) by s-GHD. These definitions generalize all Hausdorff based distances mentioned above, which can be represented by their directed distances as follows:

**HD (2):** $h(A, B) = h_{1,1}^{(p)}(A, B) = d_{11}^1$;

**PHD (4):** $h_K(A, B) = h_{K,1}^{(p)}(A, B) = d_{K1}^1$;

**CHD (5):** $h_{K,L}(A, B) = h_{K,L}^{(p)}(A, B) = d_{KL}^L$;

**MHD (6):** $h_{\text{MHD}}(A, B) = h_{1,1}^{(s)}(A, B) = \frac{1}{N_A} \sum_{i=1}^{N_A} d_{i1}^1$;

**LTS-HD (9):** $h_{\text{LTS}}(A, B) = h_{K,1}^{(s)}(A, B)$. 


We parameterize GHD replacing \( k \) and \( l \) in (11) and (12) by parameters \( \alpha \) and \( \beta \):

\[
\alpha = \frac{k - 1}{N_A}, \quad \beta = \frac{l - 1}{N_B}.
\]

Since \( 1 \leq k \leq N_A \) and \( 1 \leq l \leq N_B \) we have \( \alpha, \beta \in [0, 1) \).

We define bounded modifications of point distances:

\[
\rho^{(\tau)}(a, b) = \min\{\rho(a, b), \tau\},
\]

where \( \tau \) is a positive number. Replacing \( \rho \) with \( \rho^{(\tau)} \) in formulas (10) we introduce a new parameter \( \tau \) for GHD. So for defining a concrete p- or s-GHD, we have to choose values for the parameters \( \alpha, \beta, \rho \) and \( \tau \). Note that M-HD (7) with the function (8) coincides with MHD (6) applying \( \rho^{(\tau)} \) for point distance.

**Measuring searching effectiveness**

The effectiveness of searching methods is usually given by standard estimations of recall and precision (see M. Junker *et al.*). Let us look for a word \( W_0 \) (pattern word) in a collection of binary text images in which \( W_0 \) occurs \( N \) times. Comparing \( W_0 \) with other words in the text, a sequence of words is generated:

\[
\{W_i\}_{i=0,1,...} \quad (13)
\]
which is ordered according to a similarity measure $H$, i.e. $H(W_i, W_0) \leq H(W_j, W_0)$ for every $i < j$.

For a positive integer $n$, let $m(n) \leq n$ be the number of words among the first $n$ words of (13) that coincide with $W_0$ as words. Then recall $r(n)$ and precision $p(n)$ are defined by

$$r(n) = \frac{m(n)}{N} \quad \text{and} \quad p(n) = \frac{m(n)}{n}.$$  \hspace{1cm} (14)

$m(n)$ is nondecreasing function and the graph of $P : D \subset [0, 1] \rightarrow [0, 1]$ defined by $P(r(n)) = p(n)$ represents the effectiveness of searching methods.

**Experiments**

We define two implementations of s- and p-GHD denoted by:

- $(\alpha, \beta, \tau), p$ – the sorting algorithm for producing the word sequence (13) uses primary sort key p-GHD and secondary sort key s-GHD. This approach avoids the discontinuity of p-GHD when the words in the sequence (13) are divided into a few classes, which correspond to equal distances to the pattern.

- $(\alpha, \beta, \tau), s$ – the sorting algorithm uses primary sort key s-GHD and secondary sort key p-GHD.

In all experiments $n \in [1, 500]$. 
Bulgarian typewritten text of 333 bad quality pages is the data used in our experiments.

A word Пазарджик is a pattern word $W_0$. It occurs 231 times in the text but the number of correct segmented words Пазарджик is 200, so we set $N = 200$. 
We can see that almost 80% of words are placed at the beginning of the sequence (13) in (0.03, 0.005),s-case. The best precision 0.77 with maximum recall 0.95 is reached for (0.03, 0.005),p. The remaining parameters for all cases are \( \rho = \rho^{(\tau)}_2 \) and \( \tau = 15 \).

The best results for the word Пазарджик, obtained in our experiments for s-case and \( \rho = \rho^{(\tau)}_\infty \), are given in the next figure. We see that there is no best set of parameters – the maximum \( r(n) = 0.825 \) for \( p(n) = 1 \) is reached for \( (0.01, 0.001) \) and \( \tau = 15 \) while for \( r(n) \in [0.9, 0.975] \) the best parameters are \( (0.03, 0.005) \) and \( \tau = 19 \).
The carried out experiments are based on an old book (1884) – Bulgarian Chrestomathy, created by famous Bulgarian writers Ivan Vasov and Konstantin Velichkov. Theoretically we can find all words in the printed text which coincide with a given pattern word under the assumption that scanned images are perfect. In this instance the quality of scanned images are quite bad. Many pages have slopes in the rows, there are significant variations in gray levels, etc. There is no text version till now of this book, which might be produced using appropriate OCR software. The reasons are the quality of images and the absence of OCR software because the text contains old and obsolete Bulgarian letters. Also spelling and grammar are quite different in modern Bulgarian language.

For our experiments 200 images from about 1000 scanned pages are used. We choose a pattern word всички. It is tedious to count all words всички in all 200 pages, but we can...
estimate quite precisely their number. The best searching result gives us 114 correct words in the first 500 of the sequence (13). The total number of checked words with approximately same length is 7505 and the distribution of correct words is the reason for setting \( N = 120 \) and using this number in formulas (14).

The Figure presents the results of applying GHD for \( \alpha = 0.01, \beta = 0.001, \rho = \rho_2^{(\tau)} \) and \( \tau = 15 \). The graphics A,s and A,p are produced with the pattern word всички - s- and p-case respectively.

In the text there are two cognate words всичка and всичко. When we count as correct all three of them, setting \( N = 230 \) the obtained results are better as it can be seen in Figure, graphics B,p and B,s.
The text under investigation is Slavonic manuscript collection, “Zlatoust” (1574), 747 pages, but we consider 200 pages for the experiments. The segmentation is quite good due to the clerkly hand of the writer, and a relatively simple algorithm could separate rows and words.

The pattern word is \( 	ext{тако} \). Occasionally the same word is written as \( 	ext{тако} \). We count both words as correct retrievals. There are two more words \( 	ext{лака} \) and \( 	ext{лакъ} \) which are very similar as images but have different meanings and we do not count them. When calculating \( r(n) \), we suppose that \( N = 160 \) because there are maximum 159 correct words in the first 500 of the sequence (13), which consist of 4982 words with approximately same length.
The results show that the search process is the most successful for $\alpha = \beta = 0$ in p-case. The point distance is $\rho_2$, the parameter $\tau = 15$ for $\alpha = \beta = 0$ and $\tau = 19$ for $\alpha = 0.1$ and $\beta = 0.01$. 
Conclusions

The experiments show that the direct approach for searching words in binary text images could be applied successfully in practice. HD and its modifications are a good choice for measuring word image similarities. GHD unifies the HD approach – GHD comprises of many existing word matching methods and offers new methods by choosing various values for the parameters $\alpha$, $\beta$, $\tau$ and point distance, and processing s- or p-cases. The recommended values for $\alpha$ are in the interval $[0, 0.1]$ and for $\beta$ in $[0, 0.01]$. All three distances $\rho_2$, $\rho_1$ and $\rho_\infty$ can be used. The value of $\tau$ depends on image sizes but it must be greater than 5. There is no universal optimal parameter values for any scanned document and any searched word. The choice of good parameter values is made easier by using oriented software tool. Quite acceptable results can be achieved for $\alpha = \beta = 0$ when the image quality is relatively good.

Obtaining a word sequence for a given pattern word ordered by p- and s-GHD, using primary and secondary sort keys, gives an additional advantage in practical aspects. The experiments with Bulgarian typewritten text, printed text and manuscript confirm the possibility of wide application of our approach.


