2011 International Conference on Document Analysis and Recognition

Efficient Word Recognition Using A Pixel-Based Dissimilarity Measure

Sebastian Colutto
University of Innsbruck,
Department for Digitisation and Digital Preservation,
Innrain 52, Innsbruck, Austria
sebastian.colutto@uibk.ac.at

Basilis Gatos
Computational Intelligence Laboratory,
Institute of Informatics and Telecommunications,
National Center for Scientific Research Demokritos
GR-153 10 Agia Paraskevi, Athens, Greece
bgat@iit.demokritos.gr

***Abstract*—In this paper, we propose a word recognition methodology based on a novel size-normalization and a pixel-based image dissimilarity measure. As a first step, we apply a new size-normalization technique using baseline estimation. Starting from those size-normalized images, the difference between two word images is calculated using an image dissimilarity measure based on curvature estimation using integral invariants and a windowed Hausorff distance. We conducted several experiments comparing the new methodology with state-of-the-art techniques using ground truth data from a historical book. The experiments prove the efficiency of the proposed size normalization as well as of the overall proposed sytem.**

***Keywords*-word recognition, size normalization, distance metric**

I. Introduction

With huge amounts of documents accumulating in modern libraries, the need for digital content retrieval in such documents has become a great center of attention in the past years. While the standard approach is to digitize documents using optical character recognition methodologies, several alternative approaches for information retrieval in printed documents have been proposed for many reasons such as the high costs for standard digitization or the inability of modern OCR engines to cope with documents where severe degrations occur. An alternative approach to search a scanned document is provided via so-called keyword spotting techniques, where the user searches the document for a certain text (i.e. the keyword) rather than digitizing each character. A detailed survey to retrieval and indexing of document images in general can be found in the work of Doermann [1], or more recently an extensive survey of keyword spotting techniques was published by Murugappan et al. in [2]. There, the authors roughly divide keyword spotting techniques in methods based on character recognition on the one hand, where each single character of a given query text is recognized and searched for and word recognition techniques on the other hand, where the entire word image is used for retrieval. Our approach falls into the latter category. In particular, in this paper, we focus on a new technique for measuring the distance between two word images in a document. Several approaches in this field have been proposed in the literature, notably the popular work of Manmatha et al. [3], [4], where they use dynamic time warping with a set of features for word distance computation. Other techniques include the work of Lu et al. [5], who apply a weighted Hausdorff distance in a segmentation free template matching approach for word recognition. Gatos et al. [6] presented a combination of word image normalization and feature extraction methods for cursive handwriting word recognition. Recently, Jawahar et al. [7] have proposed word image matching methods based on learning a query specific weighted Euclidean distance.
Our technique is based on a segmentation of the document image on word level. It utilizes a novel size normalization, that uses baseline detection to optimize the size normalization procedure. Additionally, we measure the distance between two size normalized images using a combination of a windowed Hausdorff measure as proposed by Baudrier et al. in [8] and robust curvature estimation using integral invariants (cf. Manay et al. [9]). All experiments conducted show that the new size normalization in combination with the dissimilarity measure as proposed provides enhanced results when compared to other feature extraction and distance computation methods.

II. Related Work on Word Recognition Features

In this Section, we present the details of related works we compare against in the experimental results Section.

A. Zoning Features

Our first set of features to compare against were simple zoning features [10]. Zoning features first divide the image into $M \times N$ number of blocks in x and y direction. Afterwards, for each block the number of foreground pixels is counted and divided by the total number of pixels in this block which yields a feature vector of size $s = MN$ for each image.

B. Horizontal and Vertical Projection Profiles Divided Into Zones

Additionally horizontal and vertical projection profiles based on the centroid of the binary image as described in [11] are computed. First, the distance (i.e. the number of pixels) from the center of mass to the farthest pixel of

1520-5363/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDAR.2011.224

the image on the left, right, top and bottom side is measured yielding so called vertical and horizontal projection profiles. Afterwards, each of the vertical and horizontal profiles are divided into M number of blocks and the average value for each block is calculated and stored in a final feature vector of size $s = 4M$.

C. Dynamic Time Warping on Features Set

Dynamic time warping is an algorithm that efficiently computes the similarity between two sequences which can have different length. Applied to the problem of word recognition in historical documents, Rath and Manmatha have proposed a set of efficient features in [3], [4]. Those features include the upper and lower word profile, i.e. the distance from the upper and lower boundary of the word image to the closest foreground pixel. Furthermore, the number of pixel transitions (i.e. when a pixel changes from foreground to background) in each column of the image as well as the gray scale variance for all gray values in a column are also used as features.

III. Size Normalization

In order to perform the distance computation methodology presented in Section IV, we have to resize all word images to a fixed size. Starting from a binarized document image segmented on word level, all word images are normalized to this size. In Subsection III-A we first revise a default size normalization scheme against which we compare our results in our experiments, while Subsection III-B deals with the new size normalization used for word recognition based on baseline detection.

A. Default Size Normalization

The default size normalization for a word image I of size $M \times N$ to I_n of size $M^\dagger \times N^\dagger$ first calculates the centroid $(\overline{x}, \overline{y})$ of I:

$$a = \sum_{x=1,y=1}^{M,N} I(x,y), \tag{1}$$

$$b = \sum_{x=1,y=1}^{M,N} xI(x,y),\ c = \sum_{x=1,y=1}^{M,N} yI(x,y)$$

$$\overline{x} = \frac{b}{a},\ \overline{y} = \frac{c}{a}$$

Afterwards, the image is extended on each border to size $M^{'} \times N^{'}$, such that its centroid corresponds with the geometric center of the image $(x_c, y_c) = (M^{'}/2, N^{'}/2)$ and the size ratio $M^{'}/N^{'}$ of the image is the same as for the resulting size (i.e. $M^{'}/N^{'} = M^\dagger/M^\dagger$). This step ensures that a subsequent resizing step does not change the width to height ratio of the image. Finally, we resize all word images to a fixed size $M^\dagger \times N^\dagger$ using a standard bilinear interpolation scheme. An example of a size normalized word image can be seen in Figure 1.

Figure 1. Default size normalization of the word 'position'.

B. The proposed Size Normalization

At the proposed normalization scheme, we resize the word image in order to fit in a rectangular box $X_n \times Y_n$. The positioning of the word in the rectangular box is achieved by placing the baseline of the word in the center of the rectangular box. Baseline detection is accomplished using the following methodology based on horizontal projections: Let $I(x, y)$ be the binary word image array having 1s for foreground and 0s for background pixels. Also, let the rectangular bounding box of the word image has coordinates $(x_1, y_1) - (x_2, y_2)$. We first calculate the horizontal word projection LP as follows:

$$LP(y) = \sum_{x=x_1}^{x_2} I(x,y) \tag{2}$$

where $Y = y_1, \ldots, y_2$. Then, we calculate the global maximum of LP for $y = y_m$. The offset of the upper baseline y_u is estimated as follows: Starting from $y_u = y_m$ we start decreasing the value of y_u until $LP(y_u) > LP(y_m)/4$. In the same way, we calculate the offset of the lower baseline y_l: Starting from $y_l = y_m$ we start increasing the value of y_l until $LP(y_l) > LP(y_m)/4$. An example of word baseline detection is given in Figure 2. The normalized $M \times N$ word

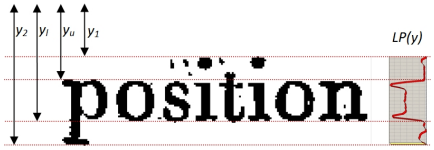

Figure 2. Example of word baseline detection.

image $I_n(x, y)$ is then calculated as follows:

$$I_n(x,y) = I(dx + e, d^\dagger y + e^\dagger) \tag{3}$$

where

$$d = \frac{(x_2 - x_1)}{M},\ e = x_1,$$

$$d^\dagger = \frac{3(y_l - y_u)}{N},\ e^\dagger = 2y_u - y_l$$

An example of the word image size normalization can be seen in Figure 3.

IV. Distance Computation

To compute distances between word images, we use an image dissimilarity measure first introduced in [12]. There,

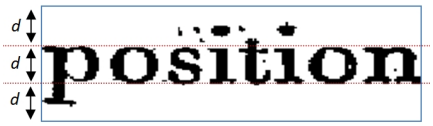

Figure 3. Proposed size normalization of the word image of Figure 2

this metric was used in an application for clustering character glyphs from historical document images. However, we show that it can also be beneficial when applied to word recognition in historical documents. The following subsections revise the metric.

A. Local Distance Map

The Local Distance Map (LDMAP), introduced by Baudrier et al. [8], is theoretically derived from the windowed Hausdorff distance, which measures the Hausdorff distance between two point sets under a predefined neighborhood W. Fixing the sliding window W locally, allows us to compute the local dissimilarity between two point sets. The LDMAP can then be computed efficiently using the distance transform DT of a point set. Given two point sets A and B of $\mathbb{R}^2$, then

$$\forall x \in \mathbb{R}^2$$
$$\text{LDMAP}(x) = |A(x) - B(x)| \max\left(DT_A(x), DT_B(x)\right) \tag{4}$$

Treating binary images as point sets, the LDMAP produces an image that measures the local dissimilarity between two binary images as illustrated for two sample words in Figure 4. In the bottom left, the absolute difference between the two word images of the top row is shown, while the lower right image shows the corresponding LDMAP for the two words. Darker areas denote higher values in this picture. Note, how the LDMAP penalizes difference-pixels for an image by their minimal distance to the border of the other image.

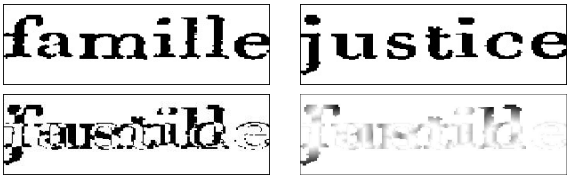

Figure 4. Illustration of the LDMAP.

B. Robust Curvature Estimation using the Local Area Integral Invariant

We use the Local Area Integral Invariant as a robust measure of curvature. Experiments have shown that this method yields superior results when dealing with noisy images, in comparison to other curvature estimation methodologies (cf. [9] or [13]). Let $B_r(p, x)$ be an indicator function on the interior of the ball of radius r around center p. Then for any radius r the corresponding local area integral invariant is defined as

$$I_C^r(p) = \int_{C_I} B_r(p, x)dx \tag{5}$$

with C and C_I denoting a curve and its interior respectively. Informally, the Local Area Integral Invariant parses the curve and measures the area of intersection of the ball with radius r with the interior of the curve. This measure can be used to estimate the curvature at a given point [9]:

$$\kappa(p) \cong \frac{2}{r} cos\left(\frac{I_r(p)}{2r^2}\right) \tag{6}$$

where $I_r(p)$ denotes the area of intersection of a ball with radius r with the interior of C and $\kappa(p)$ the curvature at point p.

Based on this definition, we introduce the Local Area Integral Invariant Map, AIMAP$_r$, which uses the Local Area Integral Invariant to estimate the curvature on the boundary of a binary image. Let I be a binary image and I_G the corresponding gradient image, i.e. a binary image indicating the edges of the source image. Furthermore, let B_r^d denote the discretized version of a ball with radius r in $\mathbb{R}^2$ i.e. a binary image of a disk, and $*$ be the discrete convolution between two images. Then the AIMAP$_r$ is defined as follows:

$$\text{AIMAP}_r = \left(\frac{2}{r} cos\left(\frac{I * B_r^d}{2r^2}\right)\right) \odot I_G + (I - I_G) \tag{7}$$

with $\odot$ being the componentwise multiplication of two images with same size. Figure 5 shows an illustration of the AIMAP$_r$ for a sample word image. There, $r = 2$ was used to compute the area integral invariant, which was experimentally determined to yield best results.

Figure 5. Illustration of the AIMAP$_r$ for a word image.

C. Combined Image Dissimilarity Measure

We combine the ideas of the LDMAP with the robust curvature estimation property of AIMAP$_r$. The rationale behind this is to add curvature estimation to the LDMAP which should lead to a more noise-robust measure of dissimilarity when noisy images are given, as mostly the case in historical documents.

The combined distance measure is defined as follows: Let I_1 and I_2 be two input images with size $M \times N$. Then the distance between the two images is defined as

$$d = \|\text{LDMAP}(I_1, I_2) \odot \max(\text{AIMAP}(I_1), \text{AIMAP}(I_2))\|_F \tag{8}$$

with $\max(.,.)$ denoting the componentwise maximum between two images, $\odot$ the componentwise multiplication and $\|.\|_F$ the Frobenius norm.

V. Experiments and Evaluation

In this section, we present the experiments conducted in order to test the proposed word recognition technique. Our evaluation is based on the comparison of the proposed methodology, i.e. the new size normalization of Section III-B and distance compuation as described in Section IV against standard feature extraction methods like zoning features and projection profiles in combination with no size normalization, the default size normalization of Section III-A and the new size normalization. Furthermore, we evaluate the quality of the results against a dynamic time warping algorithm with the features as described in [3].

A. Evaluation Results

We tested our methodology on a historical french book [14] that consists of 153 pages and 46197 words. For this book, the word segmentation as well as the ASCII ground truth were manually created. A sample page of the book is shown in Figure 6. To test the retrieval performance, we randomly selected five instances of the words 'France', 'Louis', 'famille', 'mort' and 'justice', thus yielding 25 queries in total. The total number of instances of those words are 44, 156, 47, 51 and 44 respectively. Let n_t be the total number of instances of a word in the ground truth and n_c the number of correct instances of the word in the first n_t number of retrieved instances. The retrieval performance is then calculated as $p = n_c/n_t$. An example query for the word 'France' using the proposed methodology is shown in Figure 7. There, $n_c = 40$ and $n_t = 44$, thus $p = 0.91$.

For the standard features (cf. Sections II-A and II-B), as well as the distance computation as outlined in Section IV (referred to in the tables as '*NewDist*'), we tested all queries on images resulting from standard size normalization as described in Section III-A and the newly proposed method as in Section III-B. For the size normalization, we chose to resize the images to size 300×90. For the zoning features, we set the number of blocks in horizontal and vertical direction to 30 and 9 respectively resulting in a feature vector of size 270. The same values were used for the block sizes of the horizontal and vertical projection profile features, i.e. a feature vector of size 78 was created. Regarding the dynamic time warping algorithm, we tested all queries on the original images as proposed in [3]. Furthermore, we also tested all queries on non-normalized images for the zoning and projection profile features.

The results of the experiments with the default, the new and with no size normalization are shown in Table I, II and III respectively. Each table shows the average retrieval performance p for each query word and the average overall retrieval performance over all query words. We can observe that the new size normalization yields superior results for each of the queries and that the proposed combined methodology using the pixel-based difference measure results in best retrieval perfomance for most queries and best overall performance for all methods. Note also, that the proposed scheme is able to achieve retrieval rates around 90% for each query showing its robust behaviour compared to other methods.

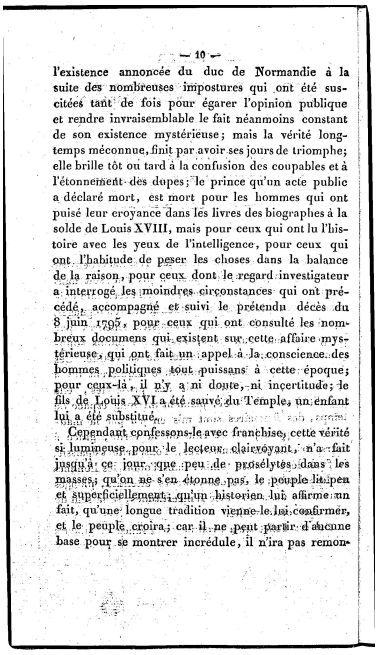

Figure 6. Sample page of the french book used for evaluation.

Table I
Query results for the different methodologies and queries using the *default* size normalization to 300×90.

	Query Word					Avg.
Method	France	Louis	famille	mort	justice	
Zoning	0.82	0.49	0.77	0.60	0.50	0.636
Proj. Prof.	0.75	0.58	0.81	0.62	0.56	0.664
NewDist	0.80	0.82	0.88	0.62	0.72	**0.768**

Table II
Query recognition results for the different methodologies and queries using the *new* size normalization to 300×90.

	Query Word					Avg.
Method	France	Louis	famille	mort	justice	
Zoning	0.93	0.93	0.75	0.85	0.65	0.822
Proj. Prof.	0.77	0.94	0.79	0.91	0.65	0.812
NewDist	0.87	0.95	0.88	0.90	0.87	**0.894**

VI. Conclusions and Future Work

We have presented a new methodology for the distance calculation in word recognition for documents. The method

Table III
QUERY RECOGNITION RESULTS FOR THE FEATURES AND THE DTW ALGORITHM AS IN [3] WITH *no* SIZE NORMALIZATION.

	Query Word					Avg.
Method	France	Louis	famille	mort	justice	
Zoning	0.42	0.25	0.42	0.45	0.68	0.444
Proj. Prof.	0.60	0.38	0.79	0.62	0.75	0.628
DTW [3]	0.78	0.90	0.91	0.87	0.80	0.852

Figure 7. Query result for the word 'France' shown at the top left. Ranked list, columnwise from top left to bottom right.

is based on a novel size normalization method and a dissimilarity measure for binary images based on the windowed Hausdorff distance and robust curvature estimation using integral invariants. Experiments on real world ground truth data have shown that the proposed approach provides overall best results when compared with standard feature extraction methods and a state-of-the-art dynamic time warping algorithm. Furthermore, it has been observed that the new size normalization enhances recognition results for all feature extraction methods and the proposed dissimilarity meausure in particular. It can also be mentioned, that we have not done any additional preprocessing on the word images, e.g. slant correction or noise removal, which is expected to improve the recogniton results of the pixel-based dissimilarity measure even further.

Future work includes experiments of the proposed approach on a larger test set as well as some improvements on the dissimilarity measure for word recognition.

ACKNOWLEDGMENT

This work has been supported by the EU 7^{th} Framework Program grant IMPACT (Ref: 215064).

REFERENCES

[1] D. Doermann, "The indexing and retrieval of document images: A survey," *Computer Vision and Image Understanding*, vol. 70, pp. 287–298, 1998.

[2] A. Murugappan, B. Ramachandran, and P. Dhavachelvan, "A survey of keyword spotting techniques for printed document images," *Artificial Intelligence Review*, vol. 35, pp. 119–136, 2011, 10.1007/s10462-010-9187-5.

[3] T. M. Rath and R. Manmatha, "Features for word spotting in historical manuscripts," in *Proceedings of the Seventh International Conference on Document Analysis and Recognition - Volume 1*, ser. ICDAR '03. Washington, DC, USA: IEEE Computer Society, 2003, pp. 218–.

[4] T. Rath and R. Manmatha, "Word image matching using dynamic time warping," in *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, vol. 2, 2003, pp. II–521 – II–527 vol.2.

[5] Y. Lu, C. L. Tan, W. Huang, and L. Fan, "An approach to word image matching based on weighted hausforff distance," vol. 0. Los Alamitos, CA, USA: IEEE Computer Society, 2001, p. 0921.

[6] B. Gatos, I. Pratikakis, A. L. Kesidis, and S. J. Perantonis, "Efficient off-line cursive handwriting word recognition," 2006, pp. 121–125.

[7] C. V. Jawahar and R. Jain, "Towards more effective distance functions for word image matching," in *Proceedings of the 9th IAPR International Workshop on Document Analysis Systems*, ser. DAS '10. New York, NY, USA: ACM, 2010, pp. 363–370.

[8] E. Baudrier, F. Nicolier, G. Millon, and S. Ruan, "Binary-image comparison with local-dissimilarity quantification," *Pattern Recogn.*, vol. 41, no. 5, pp. 1461–1478, 2008.

[9] S. Manay, D. Cremers, B.-W. Hong, A. J. Yezzi, and S. Soatto, "Integral invariants for shape matching," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 10, pp. 1602–1618, 2006.

[10] M. Bokser, "Omnidocument technologies," *Proceedings of the IEEE*, vol. 80, no. 7, pp. 1066 –1078, Jul. 1992.

[11] G. Vamvakas, B. Gatos, N. Stamatopoulos, and S. J. Perantonis, "A complete optical character recognition methodology for historical documents," in *Proceedings of the 2008 The Eighth IAPR International Workshop on Document Analysis Systems*. Washington, DC, USA: IEEE Computer Society, 2008, pp. 525–532.

[12] S. Colutto, "Introducing a new image dissimilarity measure with an application to character image clustering in degraded historical documents," in *Proceedings of the 9th IAPR International Workshop on Document Analysis Systems*, ser. DAS '10. New York, NY, USA: ACM, 2010, pp. 325–332.

[13] T. Fidler, M. Grasmair, and O. Scherzer, "Identifiability and reconstruction of shapes from integral invariants," *Inverse Problems and Imaging*, vol. 2, no. 3, pp. 341–354, 2008.

[14] *Le Dernier fils de France, ou le Duc de Normandie, fils de Louis XVI et de Marie-Antoinette, par A.* Bibliothèque nationale de France, 1838.