On the Influence of Key Point Encoding for Handwritten Word Spotting

David Fernández-Mota, Pau Riba, Alicia Fornés, Josep Lladós
Computer Vision Center, Dept de Ciències de la Computació
Universitat Autònoma de Bacelona
Email: dfernandez@cvc.uab.es

***Abstract*—In this paper we evaluate the influence of the selection of key points and the associated features in the performance of word spotting processes. In general, features can be extracted from a number of characteristic points like corners, contours, skeletons, maxima, minima, crossings, etc. A number of descriptors exist in the literature using different interest point detectors. But the intrinsic variability of handwriting vary strongly on the performance if the interest points are not stable enough. In this paper, we analyze the performance of different descriptors for local interest points. As benchmarking dataset we have used the Barcelona Marriage Database that contains handwritten records of marriages over five centuries.**

***Keywords*-Local descriptors, Interest points, Handwritten documents, Word spotting, Historical document analysis**

I. Introduction

Word spotting has become very popular for indexing and retrieval historical handwritten document images. Due to the quality of physical preservation, the writing styles, and the obsolete languages, the full transcription of such documents is extremely difficult. Therefore, huge amounts of cultural information are not accessible by contents unless users go to physically inspect them in archives and libraries, or if documents are digitized, they browse through using computer readers. Thus, to access to the information of these documents it is necessary to develop smart content-wise search procedures. Word spotting paradigms are nowadays the most effective solutions that the scientific community provides.

Word spotting is a content-based retrieval strategy where, due to the impossibility of a recognition process with enough quality, leans to a visual object detection approach. The key idea of word spotting relies upon representing word images with robust features and a subsequent classification scheme. The chosen feature space is a crucial decision. It must be representative and scalable enough to distinguish among a high number of classes (words) but invariant to the inherent variations among class instances (noise, distortion of handwriting, writing styles, etc.). On the other hand, there is a fundamental issue that not always gets the deserved relevance. It is the primitives, or interest points, on which the features are computed. Two families can be differentiated, namely appearance-based and object-based primitives.

Appearance-based methods extract descriptive features from all the image pixels in terms of the photometry. Different arrangements can be considered when analyzing the pixels, so an implicit spatial information is encoded. A typical implementation is inspired by spatial pyramid methods where the descriptors are extracted on a regular grid and different scales. Almazan et. al [6] divide the images in equal-sized cells. For each one a HOG descriptors is computed combined with an exemplar-SVM framework. Gatos et. al [7] perform a template matching of block-based images descriptors. Rothacker et. al [8] localize the descriptors on a regular grid and uniform scale. The feature vector is constructed using a dense SIFT descriptor. Almazan et. al [9] adopt the Fisher Vector (FV) representation computed over SIFT descriptors densely from the word image. Other methods use column-wise feature descriptors. Frinken et. al [10] compute global and local features in each column. Rodriguez-Serrano et. al [11] combine Marti and Bunke [12], Zoning [13] and LGH [14] features. These methods train first the models, using the information of the entire image. Once the model in trained, the images are compared and the candidates are ranked using a similarity measure, commonly a Dynamic time Warping (DTW) or Hidden Markov Model (HMM-based) similarity. Object-based methods segment local interest points from the image, and extract features on each individual object. As in other pattern recognition domain, typical interest points in images are key points like corners or crossings, edges, skeletons or regions.

Concerning features, a big variety of descriptors have been proposed in the literature on document analysis. Zhang et. al [1] reviewed the main shape representations and description techniques. The Blurred Shape Model (BSM) descriptor [2] define their interest points using the high gradient magnitude pixels. In the Shape Context descriptor proposed by Belongie et al. [3] for handwritten characters, the interest points are the pixels of the contour. Zernike moments [4] construct descriptors using a set of complex polynomials, which form a complete orthogonal set over the interior of the unit circle. Traditional Zoning methods [5] divide the image in a grid, and each cell contributes with statistics of its content to a position of a feature vector.

As stated above, the starting hypothesis of this work is that the performance of a handwritten word spotting approach does not only rely on the features but also on the interest point model over which the features are computed.

We have used as evaluation scenario historical manuscripts, in particular the Barcelona Marriage Database [24]. We have compared different segmentation strategies to extract interest points. In particular, we have extracted descriptor from foreground pixels, background pixels, key points (end points, corners and crossings), contours and skeletons. For each interest point scheme, we have computed features using several descriptors: Loci [15] and Shape Context [3]. A comparative evaluation is reported and analysed for all key point - descriptor pairs. To compute the performance we have used the Mean Average Precision metric.

The rest of this paper is organized as follows. In section II the evaluation framework is outlined. Section III describes the word normalization and the region extraction method. Sections IV and V show the methods used to compute the interest points and features. Section VI shows the experimental results. Finally, we present the conclusions in the last section of the paper.

II. Outline of the architecture

The selection of suitable key points where the descriptors are calculated is not an evident decision. Key points focus the interest of the descriptor in a local area of the image and help us to uniquely identify an image. The goal of this work is to analyze different kinds of key points for describing handwritten word images.

To assess the influence of different categories of interest point segmentation and the subsequent descriptive features, we have developed a Bag-of-Words (BoW) architecture. Thus, each key point extracted from the word image is associated with a code word. This code word comes from a code book characterizing the most relevant features after a learning process. Afterwards a descriptor consisting in a histogram of code words is used to represent the whole image word, and a classification process is performed regarding the reference descriptors.

The system consists of five stages (Fig. 1). The first step normalizes the word images. The objective is to locate the word in the centre of a template blank image, according to its centre of mass. Otherwise, two words with different lengths would have different grid sizes, resulting in different feature vectors for similar word parts. The second step of our system is a region extraction. For the sake of assessing the influence of key points in word spotting, the BoW scheme could be applied to the whole image features. However, it does not preserve spatial information. Due to this we divide word images in regions, so the features are associated to each region. Two approaches are used. The first one detects the main body area of the word. Horizontally, the word image is divided in three regions: ascenders, main body and descenders. The second one divides the image horizontally in three equal-size regions. Vertically, the word image is divided in n equal-size regions. At the end we have an irregular grid, so the associated features for each cell have a rough spacial meaning.

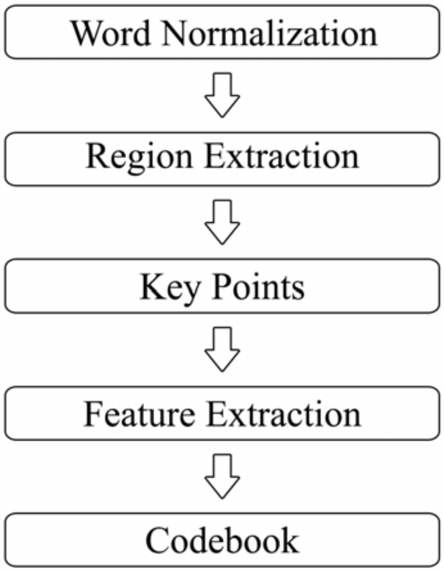


Figure 1: Flowchart of the proposed approach.

The next step consists in the extraction of the key points. Four methodologies are used to detect them in the regions of the word image. The first one uses local extrema points to select the characteristic points. The second one extracts the key points analyzing the skeleton of the word image. The third one is based in the contour of the word image. And the last ones uses the foreground and the background regions of the image. In the feature extraction step, each key point is analyzed and its feature vector is computed. A fixed-size window is located in the centre of the key point. The window is analyzed using different descriptors: Loci and Shape Context (SC).

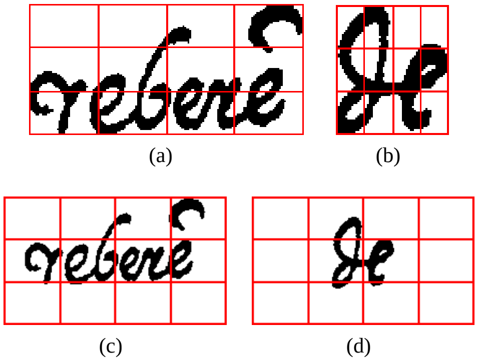
Finally, the feature vectors of a region are decoded using a model previously trained –for the SC descriptor– or using a codification algorithm –for the LOCI descriptor–. Finally all the features are stored using an histogram.

III. Word Normalization and Region Extraction

Word images are normalized regarding their center of mass. Afterwards to preserve spatial information, the images are divided into regions, so the feature vector is associated to each one.

A. Word Image Normalization

The key points used in this work are divided into horizontal regions, i.e. depending whether they appear in the top, bottom or central part of the word; and also into vertical regions. For this reason, the word length plays a crucial role in computing the similarity between words. If we divide all the word images using the same number of region/cells,



(a) (b) (c) (d)

Figure 2: Two words with the same grid size: (a) and (b). Here, the number of columns used for describing each character is different. – Words located in a blank template image: (c) and (d). Here, the number of columns used for describing each character is similar.

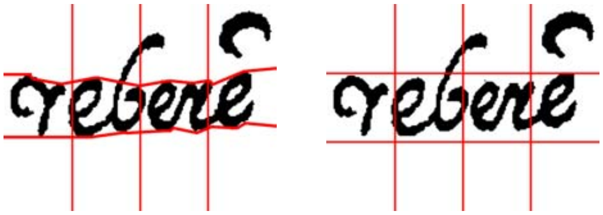
long words would have cells with a big amount of pixels in contrast of short ones with cells containing very few pixels (see Fig. 2a and 2b). The main reason is that the use of different grid sizes for each word would result in feature vectors of different sizes, and therefore, needing more complex matching techniques. So, our purpose is to describe the word images with a feature vector of the same length, so that the distance between the feature vectors can be easily performed (e.g. Euclidean distance).

For this purpose, we follow the idea proposed by Fornés et. al [16]. Every word image is located in the centre of a template blank image, according to its centroid (see Fig. 2). The advantages of the normalization before the feature extraction are: the same number of cells is used for describing the characters of each word and the feature vector of a short word (see Fig. 2d) is completely different from that long word (see Fig. 2c).

B. Region extraction

Methods based on histograms of features do not store the spatial relation of the key points. However, the use of the spatial relation can greatly increase the representation of a visual descriptor [17]. The spatial information of the key words is stored splitting the word images. The histogram of codewords is computed for each region. We have used two different configurations to split the images in horizontal regions. The first one localizes the main body area of the word and splits the word image in three regions. The second one splits the word image in three equal-sized regions using straight lines. Vertically, the word image is divided in *n* equal-sized regions.

1) Local extrema: The regions of the words are extracted using local extrema points based in the work of España-Boquera et. al [18]. The proposed method consists in automatically detecting a set of points from the image and classifying them by supervised machine learning techniques [19], [20]. The computation of local vertical extrema of foreground pixels is done in three steps: first, the contour of the image is obtained by searching positions within a column between a background and foreground pixel; second, the selected points are grouped into lines following a proximity criterion; and finally, the maxima of the upper contour and the minima of the lower contour are computed using sliding window and checking whether the central point is the maximum (or minimum) of the window.



(a) Location of the three main parts of the words (area of descenders, area of descenders and main body area)

(b) Region extraction using strigth lines.

Figure 3: Examples of the region extraction. In both cases the division in vertical is used doing straight lines.

The classification of the points in five classes (ascender, upper, lower, descender and the rest of the points) are done using two MLPs (see Fig. 4a). The first of the MLPs has two outputs with a softmax activation function to determine whether the input data is a lower baseline point or not, and the other MLP has five outputs (with also a softmax activation function) corresponding to the five classes previously described: ascender line point, upper baseline point, lower baseline point, descender line point and any other point.

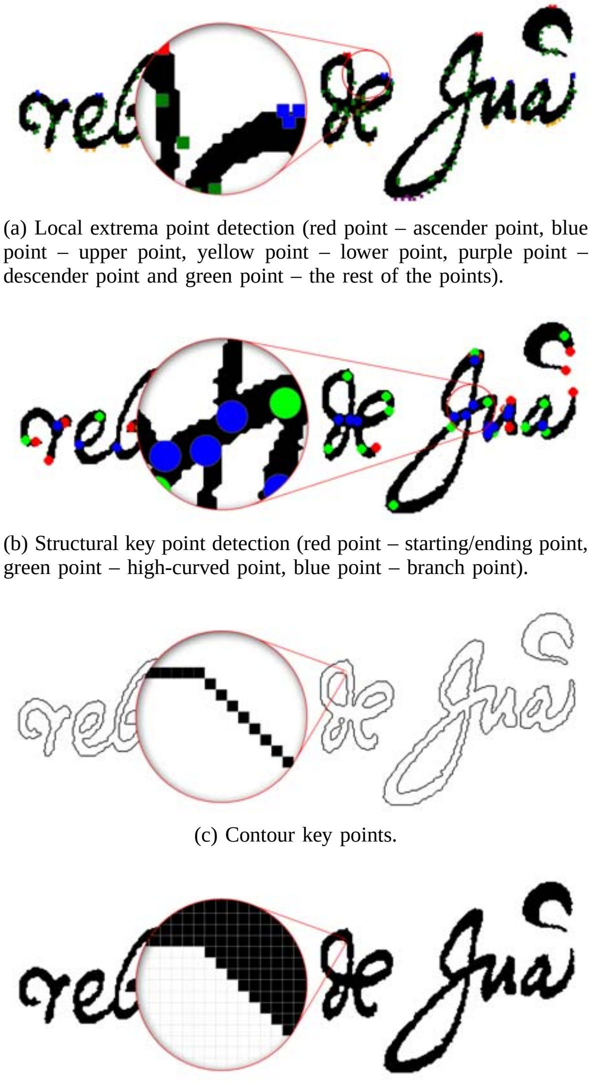
The word image is horizontally divided using the upper and the lower points (see Fig.3a). These points delimit the word image in three area: ascenders, descenders and the main body.

2) Grid cells: The word image is divided into a grid of *3 x n* equal-sized regions (see Fig. 3b). Horizontally, the word image is divided into three parts to make more comparable the two configurations used. Vertically, the word has been divided in different parts.

IV. Key points

In order to create a histogram of visual words, we first need to extract local information from the image. This local information can be extracted using characteristic points of the image, edges or regions. In our experiments we have used 4 configurations. The first two ones are based in the location of characteristic points. The third one is based in edges, and the last one in regions.

1) Local extrema: Key points are computed using the method described in Section III-B1. A feature vector is computed for each key point. Then, and in order to split the



(a) Local extrema point detection (red point – ascender point, blue point – upper point, yellow point – lower point, purple point – descender point and green point – the rest of the points).

(b) Structural key point detection (red point – starting/ending point, green point – high-curved point, blue point – branch point).

(c) Contour key points.

(d) Foreground and background key points.

Figure 4: Examples of the different interest points used.

key points in the three regions, the upper and lower points are classified as part of the main body area.

2) Skeleton-based: In this case, key points are extracted from the skeleton of the word image, as Wang et. al propose in [21].

After obtaining the skeleton of the text, structural interest points are detected. The interest points are referred to three types of points. They are respectively starting and ending points, branch points including junction and crossing points and high-curved points (see Fig. 4b). Since the skeleton is one-pixel width, for each black pixel (skeleton pixel), a 3x3 mask is applied to check the nearest 8 neighbours of the pixel. If there is only one black pixel among 8 neighbours, the reference pixel is considered as a starting or ending point. For branch points, they employed Hit-and-Miss transformation [22], a basic binary morphological operation, which is generally used to detect particular patterns in a black-and-white image. To detect high-curved points, the curvature of a point is estimated using the angle between two vectors.

3) Contour: The contour of the words can be used as key points. The subtraction of the eroded image is computed to get the contour. Every pixel of the contour is used as key point of the word image (see Fig. 4c).

4) Background and Foreground: The key points can be the foreground and background of the word image [15]. Thus, all background/foreground pixels are used as key points (see Fig. 4d).

V. Features

To test and validate the best key points scheme, we have computed several feature descriptors.

A. Loci

The *Loci* descriptor is a pseudo-structural descriptor that has been used for word spotting [15], based on the work proposed by Glucksman [23] for the classification of mixed-font alphabets. A characteristic Loci feature consists in counting the number of the intersections in eight directions: up, down, right, left and diagonals. For each interest point, and each direction, the number of intersections is counted (an intersection means a black/white transition between two consecutive pixels). Hence, each interest point generates a code, called *Locu number*, of length 8. The range of the number of the intersections is quantized and bounded in intervals. It allows a compact representation when the indexation structure is constructed because similar Locu numbers are clustered into the same bucket.

B. Shape Context

The Shape Context descriptor [3] describes the coarse distribution of the neighbouring shape with respect to a given point on the shape. For a given point in the shape contour, a coarse histogram of the relative coordinates of the remaining points is computed.

VI. experiments

In order to show the importance of the selection of suitable key points, we first describe our performance evaluation protocol in terms of the dataset, metrics and experiments.

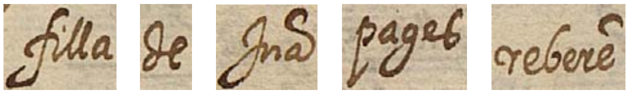
A. Dataset

For our experiments, we have used the Marriage Licenses Books conserved at the Archives of the Cathedral of Barcelona. These manuscripts, called *Llibre d'Esposalles* [24], consist of 244 books written between 1451 and 1905, and include information of approximately 550.000 marriages celebrated in over 250 parishes (Fig. 5). Each marriage record contains information about the couple, such as their names and surnames, occupations, geographical origin, parents' information, as well as the corresponding marriage fee

(a) 1618: volume 69

(b) 1729: volume 127

Figure 5: Examples of marriage licenses from different centuries.



(a) ”filla”

(b) ”de”

(c) ”Juan”

(d) ”pages”

(e) ”rebere”

Figure 6: Examples of the classes used.

that was paid (this amount depends on the social status of the family). Each book contains a list of individual marriage license records (analogous to an accounting book) during two years and was written by a different writer. Information extracted from these manuscripts is of key relevance for scholars in social sciences to study the demographical changes over five centuries.

In this work we have used the first 10 pages of the volume 69 for the experiments, and we have selected 5 of the most representative words in these documents. For each word, we have selected five random samples to compute the mean of each class. In Fig. 6 we can observe an example of each word.

B. Metrics

We use the Mean Average Precision metric to analyze the performance of the descriptors using the different key points. In the case of the Shape Context descriptor, we need to create a codebook from the word image samples. For this purpose, we have used 5 documents (1707 words). We have used a equal-sized cells to compute the key points.

C. Results

We have evaluated the different key points (described in Section IV) using two region extraction methods (see Section III). For the horizontal division, we have used two methodologies. The first one computes the three main areas of a word: main body, ascenders and descenders. The second one divides the image in 3 equal-sized regions. Vertically, the image has been divided in 3, 4 and 5 equal-sized regions. We have computed features using several descriptors (see Section V): Loci and Shape Context.

We have performed two experiments. The first one uses the local extrema points to extract the regions (Table I), and the second one uses equal-sized regions (Table II). In both cases the results are computed using different number of vertical divisions to evaluate the importance of the spatial information. Using these configurations, a comparative evaluation is computed and analyzed for all the key point – descriptor pairs.

	Loci			Shape Context		
	n=3	n=4	n=5	n=3	n=4	n=5
Skeleton Based	0.32	0.47	0.49	0.15	0.16	0.17
Local Extrema	0.39	0.56	0.54	0.50	0.60	0.66
Contour	0.45	0.73	0.61	0.56	0.50	0.51
Foreground	0.27	0.38	0.20	0.53	0.57	0.56
Background	0.77	0.77	0.78	0.57	0.58	0.59

Table I: Region extraction: *Local Extrema*.

	Loci			Shape Context		
	n=3	n=4	n=5	n=3	n=4	n=5
Skeleton Based	0.34	0.59	0.55	0.18	0.18	0.20
Local Extrema	0.32	0.51	0.52	0.47	0.44	0.53
Contour	0.31	0.55	0.55	0.53	0.53	0.69
Foreground	0.27	0.49	0.49	0.58	0.50	0.67
Background	0.73	0.76	0.74	0.60	0.60	0.67

Table II: Region extraction: *Grid Cells*.

We can observe that the number of regions is an important parameter and increasing the number of regions, the performance increases in most cases. E.g. we can see in Table II that using contours as key points the performance increase in both descriptors: from 31% to 55% and from 53% to 69%. The spatial information is more discriminant, so that key points with similar descriptors, but different spatial information, can be easily distinguished.

From the point of view of the selection of the key points, the more key points, the better the performance. The skeleton based method obtains the worst results because the computed key points are centred only in bifurcations and extrema points. However, local extrema, contour and foreground methods obtain similar results, better than the skeleton based method because the number of key points is higher and more sparse than in the skeleton based approach. The background key points outperform all the other configurations independently of the descriptor that is chosen.

VII. Conclusion

The objective of this work is to show that the performance of a handwritten word spotting approach does not only rely on the descriptor but also on the key point detection method. We have presented a comparative study showing that features computed at background key points outperform other types of key points. The main reason is that this scheme contains a more dense information, and it is more robust to the variation of strokes.

The performance of less dense key point models decreases due to the big variety of handwriting styles. For example, two feature vectors extracted from two corresponding skeleton points of two instances of the same word can vary a lot. The higher is the number of key points, the higher is the tolerance to styles.

Despite showing the importance of the key points in front of the descriptor, they are not completely independent. From the experimental results we see that there is a relation between them. Shape Context is a descriptor based in the contour of the word image, and the information of this area strongly influences the feature vector. However, Loci stores the information of the key point and its neighboring points by computing the number of crossings. Therefore, the performance of the Loci descriptor significantly increases when the key point is the background. In fact, the increase of the performance is significantly higher in Loci than in the Shape Context descriptor.

Acknowledgment

The authors thank the authors Joan Pastor and Peng Wang to provide us the key points computed with their approaches. This work has been partially supported by the Spanish projects TIN2011-24631 and TIN2012-37475-C02-02, by the EU project ERC-2010-AdG-20100407-269796.

References

[1] D. Zhang and G. Lu, "Review of shape representation and description techniques," *PR*, 2004.

[2] S. Escalera, A. Fornés, O. Pujol, P. Radeva, G. Sánchez, and J. Lladós, "Blurred Shape Model for binary and grey-level symbol recognition," *PRL*, 2009.

[3] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *PAMI*, 2002.

[4] A. Khotanzad and Y. H. Hong, "Invariant image recognition by zernike moments," *PAMI*, 1990.

[5] S. Impedovo, G. Pirlo, R. Modugno, and A. Ferrante, "Zoning methods for hand-written character recognition: An overview," in *ICFHR*, 2010.

[6] J. Almazán, A. Gordo, A. Fornés, and E. Valveny, "Efficient Exemplar Word Spotting," *BMVC*, 2012.

[7] B. Gatos and I. Pratikakis, "Segmentation-free Word Spotting in Historical Printed Documents," *ICDAR*, pp. 271–275, 2009.

[8] L. Rothacker, M. Rusinol, and G. a. Fink, "Bag-of-Features HMMs for Segmentation-Free Word Spotting in Handwritten Documents," *ICDAR*, 2013.

[9] J. Almazán and A. Gordo, "Handwritten Word Spotting with Corrected Attributes," *ICCV*, 2013.

[10] V. Frinken, A. Fischer, R. Manmatha, and H. Bunke, "A novel word spotting method based on recurrent neural networks," *PAMI*, 2012.

[11] J. Rodríguez-Serrano and F. Perronnin, "A model-based sequence similarity with application to handwritten word spotting," *PAMI*, 2012.

[12] U.-V. Marti and H. Bunke, "Using a statistical language model to improve the performance of an hmm-based cursive handwriting recognition systems," 2002.

[13] A. Vinciarelli, S. Bengio, and H. Bunke, "Offline recognition of unconstrained handwritten texts using hmms and statistical language models," *PAMI*, 2004.

[14] J. A. Rodriguez and F. Perronnin, "Local gradient histogram features for word spotting in unconstrained handwritten documents," in *ICFHR*, 2008.

[15] D. Fernández, J. Lladós, and A. Fornés, "Handwritten word spotting in old manuscript images using a pseudo-structural descriptor organized in a hash structure," in *IbPRIA*, 2011.

[16] A. Fornés, V. Frinken, A. Fischer, J. Almazán, G. Jackson, and H. Bunke, "A keyword spotting approach using blurred shape model-based descriptors," in *HIP*, 2011.

[17] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *CVPR*, 2010.

[18] S. Espana-Boquera, M. J. Castro-Bleda, J. Gorbe-Moya, and F. Zamora-Martinez, "Improving offline handwritten text recognition with hybrid hmm/ann models," *PAMI*, 2011.

[19] J. Gorbe-Moya, S. E. Boquera, F. Zamora-Martínez, and M. J. C. Bleda, "Handwritten text normalization by using local extrema classification." in *PRIS*, 2008.

[20] P. Y. Simard, D. Steinkraus, and M. Agrawala, "Ink normalization and beautification." in *ICDAR*, 2005.

[21] P. Wang, V. Eglin, C. Garcia, C. Largeron, and A. McKenna, "A comprehensive representation model for handwriting dedicated to word spotting," in *ICDAR*, 2013.

[22] D. Zhao and D. G. Daut, "Morphological hit-or-miss transformation for shape recognition," *Journal of Visual Communication and Image Representation*, 1991.

[23] H. Glucksman, "Classification of mixed-font alphabets by characteristic loci," *Proc. IEEE Comput. Conf.*, Sep. 1967.

[24] V. Romero, A. Fornés, N. Serrano, J. A. Sánchez, A. H. Toselli, V. Frinken, E. Vidal, and J. Lladós, "The ESPOSALLES database: An ancient marriage license corpus for off-line handwriting recognition," *PR*, 2013.