Historical Handwritten Text Images Word Spotting through Sliding Window HOG Features

Federico Bolelli, Guido Borghi, Costantino Grana

Dipartimento di Ingegneria "Enzo Ferrari"
Università degli Studi di Modena e Reggio Emilia
Via Vivarelli 10, Modena MO 41125, Italy
{name.surname}@unimore.it

Abstract. In this paper we present an innovative technique to semi-automatically index handwritten word images. The proposed method is based on HOG descriptors and exploits *Dynamic Time Warping* technique to compare feature vectors elaborated from single handwritten words. Our strategy is applied to a new challenging dataset extracted from Italian civil registries of the XIX century. Experimental results, compared with some previously developed word spotting strategies, confirmed that our method outperforms competitors.

Keywords: Word Spotting; Handwriting Recognition; Indexing.

1 Introduction

The transition from handwritten to digitalized historical documents establishes a great challenge, due to the huge amount of documents, the peculiarity of this kind of data, and the noise on manuscripts: generally, automatic handwriting recognizers, also called *Optical Character Recognizers* (OCRs), or standard text analyzers fail.
In this context, we develop a new word spotting technique, or rather the ability to create word collections grouped into clusters containing all instances of the same word. The creation of these clusters is based on image matching results [10]. In this way, it is possible to semi-automatically index the content of handwritten historical documents. Manual transcription and index generation is extremely expensive and time-consuming in these cases and thus not always feasible for voluminous manuscripts.

In this paper we propose a method to extract features from historical document words and to match them exploiting the *Dynamic Time Warping* (DTW) technique, which compares and aligns feature vectors elaborated from single handwritten words. We collect a new dataset that is publicly available and it is acquired through a previously developed system of image dewarping [1]. This system starts from a curled page, usually taken by a digital scanner or digital camera, and outputs an image constituted only of horizontal straight text lines, without any distortion due to perspective, lenses and page warping. In particular,

through this approach, a great amount of documents from Italian civil registries of the XIX century are available for our scope.

The paper is organized as follows. Section 2 presents an overall description of related literature works. In Section 3, the proposed method is detailed. The proposed dataset is described in Section 4. Section 5 reports experimental results. Finally, in Section 6 conclusions are drawn.

2 Related Work

The original idea of word spotting for handwritten manuscripts was initially presented in [8] and [9]. In these works, matching techniques and pruning methods are described: given a word's bounding box, unlikely matches are quickly discarded and similar words are clustered.
Generally, word spotting methods can be divided in two main classes: *line-segmentation* and *word-segmentation* based approaches.

Line-segmentation based methods rely on the hypothesis that each line in the document is separated and word segmentation techniques are not strictly required. Terasawa *et al.* [14,13] presented a word spotting method based on line segmentation, sliding window, continuous dynamic programming and a gradient-distribution-based feature with overlapping normalization and redundant expressions, also known as "slit style HOG features". In [6] a line-oriented process is applied to avoid the problem of segmenting cursive script into individual words. This approach exploits pattern matching techniques and dynamic programming algorithms. The presented system is tested on old Spanish manuscripts, showing a high recognition rate. Even the adoption of a number of heuristic to limit the search along document lines, this approach is expensive since words have to be searched for every possible position. Besides, DTW is separately applied on each feature vector and results are heuristically merged, producing different alignment for the same word-line pair.

On the other hand, *word-segmentation* approaches are based on the hypothesis that each word in the document images is separately clipped. A word-by-word mapping between a scanned document and a manual transcript is proposed in [15]: in this way, it is possible to exactly locate words in document pages. This method relies on a OCR used as a recognizer for multiple word segmentation hypothesis generated for each line of the document. Results shown that OCR is not a useful and feasible solution for historical manuscript recognition. In [11] a local descriptor, inspired by the SIFT [7] key-point descriptor, is proposed. Significant improvements are achieved exploiting two different word spotting systems, based on the well-known *Hidden Markov Models* and DTW.
In [10] a range of features suitable for DTW has been analyzed: this work is described in detail because we use it as a touchstone. Speed and precision have achieved as result of combining different text features which are extracted from pre-processed rectangular word images and that do not contain ascenders from other words. Moreover, inter-word variations such as *skew* and *slant* angles are detected and normalized. Investigated features include projection profile, partial

Fig. 1: Example of pre-processing steps applied to an handwritten word image. (a) is the raw input grayscale word image, (b) the result after binarization process, (c) the graphical output of the RLSA algorithm. Connected components are then labeled (d) and the bounding box of the biggest component is extracted (e) and (f). In (g) is reported the output of *Canny* algorithm applied on (e).

projection profile, upper and lower word profile, background to ink transitions, gray scale variance, and feature sets containing horizontal and vertical partial derivatives applied through a Gaussian kernel. Best performance in terms of average precision are achieved by the combination of projection profile, upper and lower profile and ink transitions. In order to compare this strategy with our proposal, we produce an implementation of this algorithm, maintaining all details described in the corresponding paper.

3 Proposed Word Spotting Method

The method proposed in this paper is *word-oriented*, thus we describe it starting from single word image as the one reported in Figure 1a (see Section 4 for extraction details). Before proceeding with feature extraction all word images are pre-processed as described in the following section.

3.1 Word Image Pre-processing

All input images are binarized through an adaptive threshold [12] which deal with the light changes that occur in the original manuscripts (Figure 1b). Then, we exploit the horizontal *Run Length Smoothing Algorithm* (RLSA) [16] to ensure that all pixels belonging to the word contained in the binary image are connected (Figure 1c). The threshold used for RLSA is equal to text height that is calculated as described in [4]. Thanks to the *Connected Components Labeling*

(CCL) [5] we are able to extract the word using the bounding box of the biggest component (see Figure 1d, 1e and 1f for instance). We aim, through the combination of RLSA and CCL algorithms, to filter out all the graphical contents that do not belong to the handwritten word which are represented by remainders of other words in the original document. After these steps, it is possible to remove background from the image.
Moreover, images are vertically and horizontally resized to a fixed window of 352 $\times$ 90 pixels. This operation could be viewed as a normalization of handwritten word's width and height, and it is also a fundamental step for next elaborations. Finally, *Canny* [2] algorithm is applied on the binarized and resized image in order to make our algorithm invariant to ink thickness.
Graphical result is reported in Figure 1g.

3.2 Feature Extraction and Word Matching

In order to compare different word images through DTW, HOG descriptors [3] are computed follow a sliding windows approach. According to [14] we divide each input image in windows of fixed size 16 $\times$ 90. Windows are then split into blocks of 4 $\times$ 2 cells each composed by 4 $\times$ 4 pixels (see Figure 2b). Finally, 12 bins of the signed gradient histogram are used in orientation binning. Signed gradient produces better results in this scenario because it is generally not possible to have some characters brighter and some other darker than the background mixed in the same manuscript.
The defined block has the same width of the window so no horizontal overlapping is allowed during HOG features extraction, instead, they are vertically overlapped with a stride of two pixels for a total of 4032 descriptors per window. In our experiments we test the proposed approach using overlapped windows in horizontal direction with different strides.

As said before, DTW is exploited to compute and align the similarity distance between two given word images. The Dynamic Time Warping matching algorithm is based on the recurrence equation

$$DTW(i,j) = min \begin{Bmatrix} DTW(i-1,j) \\ DTW(i,j) \\ DTW(i,j-1) \end{Bmatrix} + d(i,j) \tag{1}$$

where $d(i,j)$ is the distance between the $i-th$ and $j-th$ feature vectors (respectively called x and y and both of length N) of the two images to match:

$$d(i,j) = \sum_{k=1}^{N} |x_{ik} - y_{jk}| \tag{2}$$

In Figure 2c is reported an example of DTW distance matrix calculated with Formula 1. In this example, HOG feature vectors are obtained with window stride equal to window width (*i.e.* non overlapped windows). The green path, usually called *warping path*, represents the optimal match between the two words. This

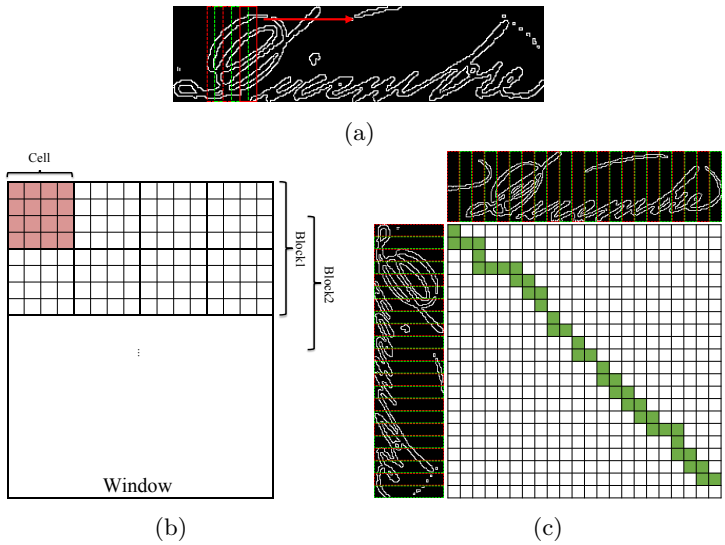

Fig. 2: (a) example of sliding window on a XDOCS's word image for HOG descriptors calculation; (b) details of the adopted windows; (c) example of DTW matrix obtained for two XDOCS's word images.

approach let us to determine a measure of word similarity independently from certain non-linear variations in the time dimension.

4 Dataset

As mentioned above, we collect a new challenging dataset. The dataset consists of a collection of handwritten month names extracted from Italian civil registries of the XIX century. We extract word images employing a template approach: given a rectified image of a whole page of the historical document, we directly extract month names placed in fixed position of the page. In this way, we can automatically collect a number of high quality and easy to annotate samples. Specifically, the obtained dataset consists of around 1200 words and all 12 months are available. Moreover, the variety of handwritten words is guaranteed by three different official state writers. The dataset is publicly available [1].

The dataset creation approach relies on the assumption that rectified pages are available. These are obtained by the use of the dewarping technique described in [1]. The entire pipeline could be summarized in these steps:

– *Image pre-processing*: in this step document and page noise is filtered out; this is mainly due to the digitization process and to the intrinsic nature of the original images;

[1] http://imagelab.ing.unimore.it/XDOCS

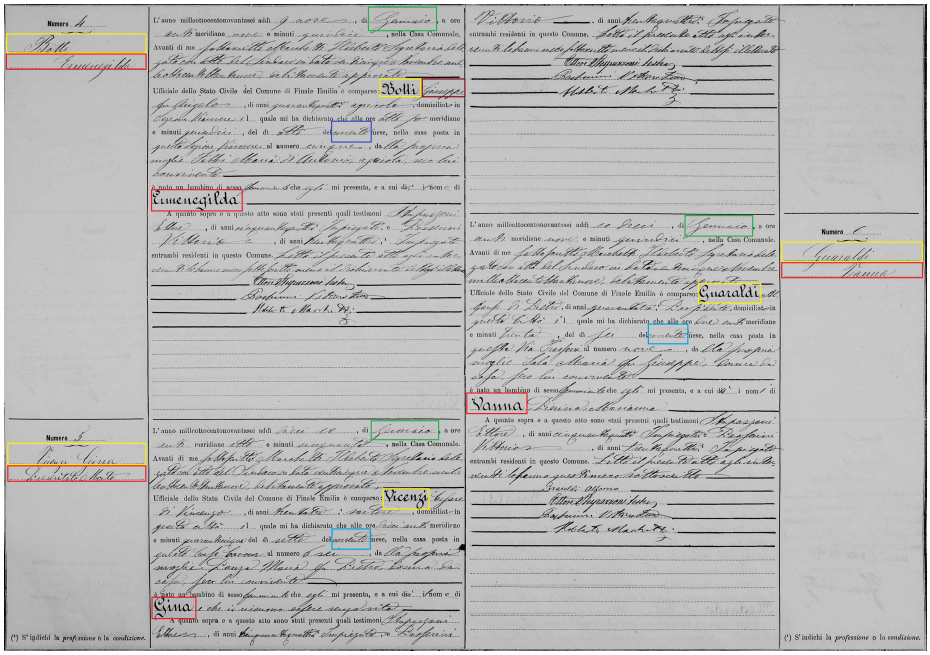

Fig. 3: Example of dewarped document and word extraction leaded by template. A template approach for word extraction is possible due to the prearranged structure in these particular historical registries.

- *Projection extraction*: this step aims to find the curved surface projection represented by two almost vertical straight lines and by two third degree polynomial curves surrounding the document page. This is required by the implemented dewarping method.
- *Dewarping phase*: this step is the core of the image rectification and dataset creation phases. During this phase, the projection of the curved surface is mapped into a rectangular normalized area.

At the end of this process, input images are correctly rectified and they do not suffer of any distortion effects. The result is depicted in Figure 3 where colored bounding boxes show word of interest automatically extracted by image coordinates. In the following we will refer to the described dataset as XDOCS dataset.

5 Experimental Results

We test our system on the XDOCS dataset divided into three different group, one for each handwriting style. We refer to each group with the name of the municipality from which original documents belong, *i.e. Vignola*, *Carpi*, and

Formigine. This approach let us to perform both *intra* and *inter* dataset evaluation.
Following a common practice for word spotting task, we exploit the *Mean Averages Precision* (MAP) with *cut-off* at $C = \{5, 10, 15\}$ (*i.e.* MAP@5, MAP@10, and MAP@15) to evaluate and compare different algorithms. Given a couple of datasets the first (with a word elements) is used to create queries that are then performed on the second one (with b word elements). For each query the average precision is calculated using the following formula:

$$ap@n = \frac{\sum_{k=1}^{n} P(k)}{\min(m, n)} \tag{3}$$

		Vignola	*Carpi*	*Formig.*
Vignola	MAP@05	0.528	0.100	0.181
	MAP@10	0.380	0.086	0.144
	MAP@15	0.306	0.093	0.132
	CMF	75.25%	17.82%	26.73%
Carpi	MAP@05	0.135	0.466	0.095
	MAP@10	0.101	0.434	0.078
	MAP@15	0.079	0.414	0.072
	CMF	14.53%	63.25%	15.38%
Formig.	MAP@05	0.192	0.127	0.644
	MAP@10	0.156	0.114	0.541
	MAP@15	0.135	0.121	0.476
	CMF	24.69%	19.25%	77.82%

(a) Our - 16 pixels stride.

		Vignola	*Carpi*	*Formig.*
Vignola	MAP@05	0.634	0.108	0.206
	MAP@10	0.465	0.098	0.168
	MAP@15	0.376	0.101	0.156
	CMF	83.17%	18.32%	25.25%
Carpi	MAP@05	0.145	0.534	0.112
	MAP@10	0.110	0.489	0.093
	MAP@15	0.086	0.485	0.086
	CMF	17.10%	66.67%	17.95%
Formig.	MAP@05	0.268	0.150	0.775
	MAP@10	0.209	0.133	0.662
	MAP@15	0.173	0.138	0.582
	CMF	36.82%	21.34%	89.96%

(b) Our - 8 pixels stride.

		Vignola	*Carpi*	*Formig.*
Vignola	MAP@05	0.665	0.102	0.222
	MAP@10	0.493	0.093	0.189
	MAP@15	0.400	0.098	0.170
	CMF	87.13%	14.85%	27.22%
Carpi	MAP@05	0.159	0.578	0.125
	MAP@10	0.117	0.536	0.101
	MAP@15	0.091	0.527	0.096
	CMF	19.66%	73.50%	17.95%
Formig.	MAP@05	0.309	0.177	0.823
	MAP@10	0.235	0.152	0.708
	MAP@15	0.194	0.153	0.621
	CMF	40.59%	26.77%	94.14%

(c) Our - 2 pixels stride.

		Vignola	*Carpi*	*Formig.*
Vignola	MAP@05	0.468	0.042	0.077
	MAP@10	0.347	0.034	0.057
	MAP@15	0.276	0.028	0.050
	CMF	68.32%	9.90%	13.37%
Carpi	MAP@05	0.086	0.445	0.087
	MAP@10	0.060	0.411	0.067
	MAP@15	0.050	0.382	0.058
	CMF	13.78%	51.70%	15.34%
Formig.	MAP@05	0.097	0.053	0.557
	MAP@10	0.071	0.045	0.413
	MAP@15	0.060	0.042	0.342
	CMF	19.25%	9.21%	80.33%

(d) Rath *et al.* [10].

Table 1: Results (on the XDOCS dataset) of the proposed method with different window strides (a), (b), and (c) and of the algorithm described by Rath *et al.* (d).

In Eq. 3, $P(k)$ is the precision at cut-off k in the item list, *i.e.*, the ratio of the number of correct word matches, up to position k, over the number k; n is the cut-off chosen from C and m is the total number of word images that match with the query in the second dataset.

Once the $ap@n$ is calculated for every query the $MAP@n$ is given by Eq. 4.

$$MAP@n = \frac{\sum_{i=1}^{Q} ap@n_i}{N} \tag{4}$$

where $Q = a$ is the number of queries and $ap@n_i$ is the average precision for the i-*th* query.

Moreover, an additional metric is included in our evaluation: for every test among two different datasets we store the number of queries that return a correct match as first. This value is reported in tables with the acronyms CMF (Correct Match First) and in percentage with respect off the total number of queries.

We compare our method with the one described by Rath *et al.* in [10] that is one of the state-of-the-art algorithms for word spotting task. This method is chosen because it uses a comparable approach based on DTW which exploits different features to perform matching between different words. As we said in Section 2, we implemented Rath's algorithm because, from our knowledge, a public implementation is not available. According to [10], we included in our code only the combination of features which achieve better performance in the original paper.

In Table 1a, 1b, and 1c the performance of the proposed method, using a stride ranging from 2 to 16 for HOG windows, is presented. The MAP increase when the stride size decrease and as consequence the best results are obtained with stride 2. Unfortunately, the computational cost of the process increase with lower strides. According to the application in which word spotting has to be applied, a compromise between accuracy and computation time can be selected. Even thought HOG descriptors can be calculated off-line during words extraction process, the DTW suffers with very long feature vectors such as the one obtained by HOG.

On the other hand, in Table 1d results achieved by our designed competitors on the same datasets are reported. As depicted, our method achieves better accuracy in all tests also with 8 pixels stride. Moreover, CMF reveals that our proposal can be used as enabling technology for real word applications.

In Figure 4 average query execution time required by the experimented word spotting algorithms is reported. All average times are computed considering the entire XDOCS dataset, thus the search space counts almost 1200 word images. No code optimizations are involved neither in the word spotting algorithm nor in the DTW implementation, so Figure 4 serve only as comparison between time performance of the described algorithms.

6 Conclusions

In this paper two main contributions are described. Firstly, a new and challenging dataset of handwritten historical documents from Italian civil registries is

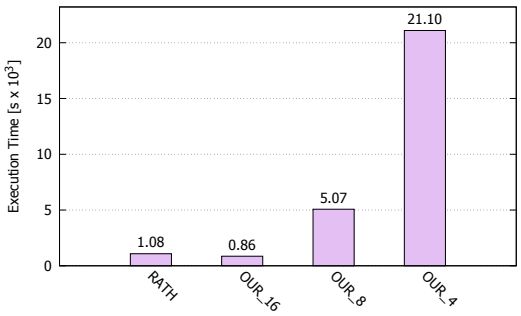

Fig. 4: Average query execution time for the algorithm by Rath *et al.* [10] and our proposal with different window strides (*i.e.* 16, 8, 4). The time required by windows stride 2 (139.2×10^3 s) is not reported to facilitate the readability of the chart. Query search space is composed by almost 1200 samples (*i.e.* the entire XDOCS dataset).

publicly released. Secondly, a novel method to tackle the problem of word spotting task is presented: it is based on HOG descriptors and exploits the Dynamic Time Warping technique to compare feature vectors elaborated from single handwritten words. The system is able to achieve a good accuracy in terms of MAP and overcomes a literature competitor.

References

1. Bolelli, F.: Indexing of historical document images: Ad hoc dewarping technique for handwritten text. In: 13th Italian Research Conference on Digital Libraries. IRCDL (February 2017)
2. Canny, J.: A computational approach to edge detection. IEEE Transactions on pattern analysis and machine intelligence (6), 679–698 (1986)
3. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. vol. 1, pp. 886–893. IEEE (2005)
4. Gatos, B., Pratikakis, I., Ntirogiannis, K.: Segmentation based recovery of arbitrarily warped document images. In: Ninth International Conference on Document Analysis and Recognition (ICDAR 2007). vol. 2, pp. 989–993. IEEE (2007)
5. Grana, C., Baraldi, L., Bolelli, F.: Optimized connected components labeling with pixel prediction. In: International Conference on Advanced Concepts for Intelligent Vision Systems. pp. 431–440. Springer (2016)
6. Kolcz, A., Alspector, J., Augusteijn, M., Carlson, R., Popescu, G.V.: A line-oriented approach to word spotting in handwritten documents. Pattern Analysis & Applications 3(2), 153–168 (2000)
7. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International journal of computer vision 60(2), 91–110 (2004)
8. Manmatha, R., Croft, W.: Word spotting: Indexing handwritten archives. Intelligent Multimedia Information Retrieval Collection pp. 43–64 (1997)

9. Manmatha, R., Han, C., Riseman, E.M., Croft, W.B.: Indexing handwriting using word matching. In: Proceedings of the first ACM international conference on Digital libraries. pp. 151–159. ACM (1996)
10. Rath, T.M., Manmatha, R.: Features for word spotting in historical manuscripts. In: Document Analysis and Recognition, 2003. Proceedings. Seventh International Conference on. pp. 218–222. IEEE (2003)
11. Rodriguez, J.A., Perronnin, F.: Local gradient histogram features for word spotting in unconstrained handwritten documents. Proc. 1st ICFHR pp. 7–12 (2008)
12. Sauvola, J., Pietikäinen, M.: Adaptive document image binarization. Pattern recognition 33(2), 225–236 (2000)
13. Terasawa, K., Nagasaki, T., Kawashima, T.: Eigenspace method for text retrieval in historical document images. In: Document Analysis and Recognition, 2005. Proceedings. Eighth International Conference on. pp. 437–441. IEEE (2005)
14. Terasawa, K., Tanaka, Y.: Slit style hog feature for document image word spotting. In: Document Analysis and Recognition, 2009. ICDAR'09. 10th International Conference on. pp. 116–120. IEEE (2009)
15. Tomai, C.I., Zhang, B., Govindaraju, V.: Transcript mapping for historic handwritten document images. In: Frontiers in Handwriting Recognition, 2002. Proceedings. Eighth International Workshop on. pp. 413–418. IEEE (2002)
16. Wahl, F.M., Wong, K.Y., Casey, R.G.: Block segmentation and text extraction in mixed text/image documents. Computer graphics and image processing 20(4), 375–390 (1982)