BH2M: the Barcelona Historical Handwritten Marriages database

David Fernández-Mota, Jon Almazán, Núria Cirera, Alicia Fornés & Josep Lladós
Computer Vision Center – Dept. Ciències de la Computació
Universitat Autònoma de Barcelona
08193 Bellaterra (Cerdanyola) Barcelona, Spain
dfernandez@cvc.uab.es

***Abstract*—This paper presents an image database of historical handwritten marriages records stored in the archives of Barcelona cathedral, and the corresponding meta-data addressed to evaluate the performance of document analysis algorithms. The contribution of this paper is twofold. First, it presents a complete ground truth which covers the whole pipeline of handwriting recognition research, from layout analysis to recognition and understanding. Second, it is the first dataset in the emerging area of genealogical document analysis, where documents are manuscripts pseudo-structured with specific lexicons and the interest is beyond pure transcriptions but context dependent.**

I. Introduction

Digitization of historical manuscripts is a priority for archives and libraries worldwide. This process aims first the digital preservation of the documents, but it raises another challenge, the access to contents. The document analysis community has been very prolific during the last decade in the release of tools for (semi)automatic recognition and annotation of historical manuscripts, as services to scholars in social sciences and humanities. Layout analysis [1], [2], including line [3], [4] and word segmentation [5], [6], and recognition, including transcription [7], [8], [9] and word spotting [10], [11], [12], [13], are the most usual tasks of a knowledge extraction process. Whatever is the task, to ensure robust methods, benchmarking databases are needed.

A Ground Truth (GT) designed to validate the interpretation of historical manuscripts has to incorporate several difficulties that hinder the recognition of manuscripts. First, the intrinsic physical effects of degradation over time that result in show-trough, bleed-through, stains, holes, etc. This phenomenon is easily guaranteed if the images come from real books. Second, the use of archaic languages, which requires specific lexicons in the GT to train classifiers. Third, the semantics of the contents themselves, which is probably the most difficulty in a GT generation because it involves the use of contextual knowledge (a palaeographer interprets historical manuscripts using the knowledge of the time and theme context). Hence, it is preferable to have domain specific datasets than generic ones with complex additional meta-data.

A number of reference historical manuscript datasets are being used by the document analysis community. In historical handwriting recognition and word spotting, a key reference is the *George Washington* dataset [14]. It is written in English language and contains 20 pages from a single writer. *Parzival* [15] is a multi-writer historical database and contains 47 pages written in medieval German language. *Rodrigo* [16] is a single writer database written in Old Spanish and contains 853 pages. Databases containing modern handwriting are also commonly used. Among them, one example is the *IAM* database [17]. It is a multi-writer database which contains instances of handwritten English text and it consists of 1,539 pages. The *CASIA* database [18] consists of a dataset written in Chinese and another one in English. It contains 1,074 handwritten texts in on-line format. The *IFNENIT* [19] is a multi writer database of Arabic handwritten texts consists of 2,200 images.

(a) Index from 1617 (b) License from 1617 (c) License from 1860

Fig. 1: *Llibre d'esposalles* (Archive of Barcelona Cathedral)

Among the different categories of historical documents, there is a growing interest in the analysis of census, birth, marriage, or death records. A number of crowdsourcing campaigns exist to transcribe information from such types of documents[1], and relevant works have been published [20], [21]. Search centered at people is very important in historical research, including family history and genealogical research. Queries about a person and his/her connections to other people allow focusing the search to get a picture of a historical context: a person's life, an event, a location at some time period. In this scenario, the challenge is not the transcription, but the understanding of the documents. This allows advanced tasks such as intelligent information extraction, summarization or knowledge discovery [22], [23].

Although such a big effort among the research community for the analysis and interpretation of digitized parish and civil records, there is a lack of specialized GT. Therefore, the first contribution of this paper is a complete database consisting in 174 images of manuscripts from the 17th century written

[1] http://familysearch.org/, http://www.ancestry.com/

over two years[2]. The meta-data provided allows to test the performance of the complete interpretation pipeline, from the layout analysis to the recognition. The second contribution is that, to the best of our knowledge, it is the first database in the area of genealogy research. Hence, the contents of these documents can be appropriately mined and exploited by means of advanced artificial intelligence and machine learning techniques. This is the reason why this database is not only a benchmark for the document analysis community. For many other disciplines it is also an everlasting source of information, that will be a legacy to the following generations.

The rest of the paper is organized as follows. Section II describes the marriage licenses books collection, the main difficulties and the possible uses. Section III presents the baseline experimental use cases. Finally, Section IV draws the conclusion and shows directions for future improvements.

II. The *BH2M* books collection

In the 15th century, a centralized register of marriages called *Llibres d'Esposalles* was created in Barcelona. Its purpose was to record all the marriages occurred in Barcelona and surroundings, as well as keep record of the *marriage fee* that was paid. It is nowadays conserved at the Archives of the Barcelona Cathedral and comprises 244 books with information on approximately 550,000 marriages held between 1451 and 1905 in over 250 parishes. Each book was written by a different writer and contains information of the marriages during two years.

Books in the collection consist of two consecutive sections. The first section is an index with all the husbands' surname on the volume, and the page number where they appear (Fig. 1a). The second section is the marriage licenses (Fig. 1b and 1c). One can clearly appreciate the continuity of the layout during the centuries along the books, but with significant differences in the handwriting styles.

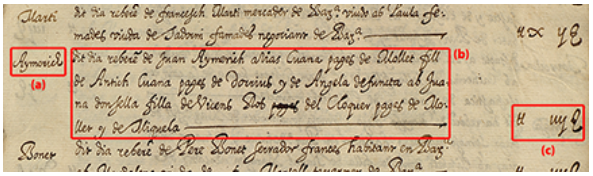

Fig. 2: Structure of the documents: (a) husband's surname, (b) license, (c) fee of the wedding.

Marriage licenses have also a structured layout (Fig. 2), which divides the document in three well-defined parts, or columns: the *husband's surname*, the *license* and the *marriage fee* column. An important characteristic of these documents is that all the licenses present a quite regular structure that can be represented by a syntactic model. The BH2M database has been generated with a crow-sourcing paradigm under the EU ERC project *Five Centuries of Marriages (5CofM)*. The database consists of two parts. First the social science view that contains the transcribed marriage licences with normalized names, and second, the GT for document analysis and recognition tasks which is presented in this paper. From the same source of images, a GT specifically designed for handwriting recognition was presented in [8]. In this paper a more generic meta-data structure is presented as a result of the *5CofM* project, covering the whole pipeline of document analysis and interpretation.

[2]The database is available at http://dag.cvc.uab.es/5cofm-ground-truth

Fig. 3: Examples of difficulties: (a) a graphical element voiding a license, (b) cross-out words, (c) a license with a special drawing, (d) the words *Jua-na* and *Mo-ller* are split in two lines, (e) the word *sastre* is written between two lines, (f) Abbreviation of *Juana* using the special symbol ˜, (g) Abbreviations of Barcelona (Bar^a) written with different shape and (h) presence of touching and horizontally-overlapped components.

A. Difficulties/Challenges

The analysis of historical handwritten documents is not a trivial process due the intrinsic difficulties of this kind of documents (Fig. 3). The physical lifetime degradation of the original documents, related to the frequent handling and careless storage, produces holes, spots, broken strokes, ink bleed, show-through, winkles, etc. In addition, during the scanning process non stationary noise can be introduced due to illumination changes, low contrast, warping, etc. These difficulties influence in the performance of tasks like word spotting, line and word segmentation, text retrieval or text alignment.

Besides these general difficulties, the characteristics of the handwriting and the configuration of the text lines may provoke additional difficulties. Curvilinear baselines due to the non-straight pen movement, lines of crowded writing styles, the presence of touching and horizontally-overlapped components [24] can introduce confusion in the recognition of the physical structure. These documents do not present a standard nomenclature, with different word spelling, special symbols, abbreviations, cross-outs, comments between lines and other recognition challenges. In addition, new words are constantly appearing along their pages.

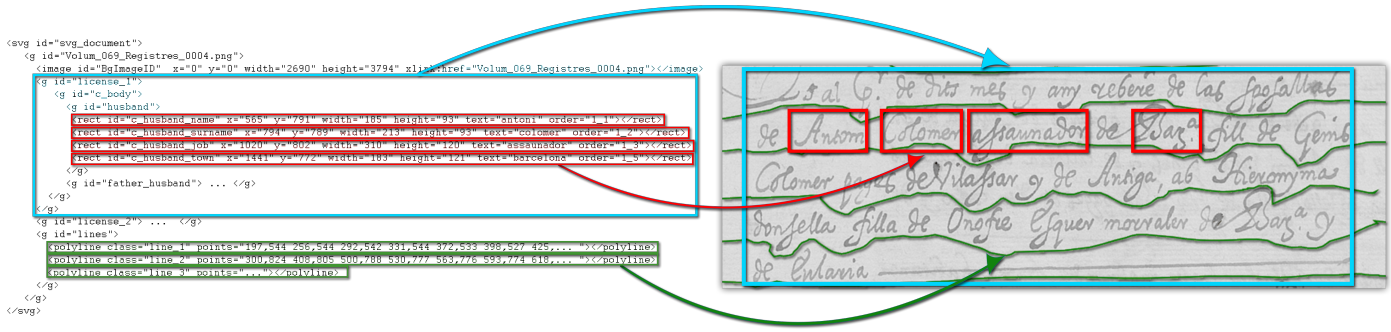

Fig. 4: A simplified example of a XML file structure containing the information of the documents. We can observe how the information is stored in the XML file and the correspondence of this information in a graphical representation.

B. Marriage Licenses

The main contribution of this paper is a publicly available database from the collection of Barcelona marriage license books. The dataset presented here corresponds to the volume 69, which contains 174 handwritten pages. This database has been compiled using the marriage licenses. This book was written between 1617 and 1619 by a single writer in old Catalan. Table I shows the figures of this volume.

Generally speaking, the database consists of annotated images. The annotations consist of an XML hierarchical structure (from individual words to blocks of text). The minimum unit of information is a bounding box of an individual word, with the corresponding transcription. Additional attributes like line, register number, or category, are associated to words. This representation allows to easily retrieve a GT from the meta-data adapted to the tasks to be tested (segmentation, word spotting, handwriting recognition, etc.). Three levels in the XML meta-data associated to images can be identified. The first one is designed to evaluate tasks for layout analysis, the second one for text transcription, and the third one for context dependent interpretation. Let us further describe these three levels of meta-data.

1) Layout structure: Document structure and layout analysis is the first processing phase applied to each page image in order to decompose it into regions. Here, each page is consist of different physical blocks: text blocks, paragraph, lines and words. The top layer correspond to the text blocks of the page. A document has three text blocks: left block or *husband's surname*, right block or *fee* and central block or *license*. This database has been created using the *license* text block. The second level corresponds to segmented lines. An accurate segmentation of the text lines is provided, including the ascenders and descenders of the text line (see Fig. 4).

The third level consists of text words (Fig. 4 – red boxes). The attributes stored for each word are the following:

- The ID of the page where the word is.
- The bounding box coordinates of the word.
- The text block where the word is located (in case the word is inside a text block).
- The text line where the word is located.
- The appearance order in the text line.

Number of:	Train	Validation	Test	Total
Pages	100	34	40	174
Licenses	998	339	403	1,740
Lines	3,132	1,065	1,301	5,498
Words	32,416	11,089	13,140	56,645
Unique word classes	3,060	1,535	1,757	3,360
Words that appear 1 time	1,831	942	1,100	1,710
Words that appear 2 times	397	213	234	552
Words that appear >2 times	832	380	423	1,098
Out of Vocabulary (OOV) words	–	594	1,082	–

TABLE I: *BH2M*: Figures.

- The word also stores information about special cases. For example, if the word is part of a title, a comment, a cross-out word, etc.

The layout meta-data allows to evaluate algorithms for layout segmentation, from individual words to the text blocks. Depending on the granularity level to evaluate (words, lines, blocks) practitioners must select to corresponding objects in the XML files.

2) Transcription: The handwritten text was literally transcribed by volunteers using a crowd-sourcing platform [25]. Given the complexity of the handwriting style due to many subtle spelling variants and the language itself, a posterior revision by experts demographers was performed to ensure its correctness. Along the 1,760 licenses, there are more than 56,000 different handwritten words which later conform 3,360 unique word classes. Table I summarizes some basic figures of the *BH2M* text transcriptions. From the text level document analysis point of view, the presented database is a suitable and syntactically enriched benchmark for various tasks such as word spotting or handwriting recognition.

Word transcriptions follow certain rules in order to unequivocally codify the text: spelling mistakes are not corrected, abbreviations are not expanded; superscript characters are denoted by the precedence of the upper symbol " ^ " and bounded by brackets (e.g. Barn^(a)); superscript of word ending character m (an upper stroke over the last characters of the word) is denoted by symbol $ (e.g. rebere$). The complete list of rules transcriptions can be found at the website.

3) Semantic information: In addition to word transcription and location, the atomic units of this database are also labelled with meta-text information. Next, we describe the mentioned semantic information.

Each marriage license has the purpose of accounting for a prospective marriage. Hence the licenses contain similar information and structure, albeit the structure can vary over centuries. The general sub-parts of a license (*L*) are, in appearance order, the date-related (*D*) information, the husband-related (*H*) information and the wife-related (*W*) information. These three sub-parts are joined by keywords as follows: date (*DI*) and husband parts are connected by *rebere de* (*we received from* in old Catalan), husband and wife (*HI* and *WI*) information are connected by *ab* (abbreviation of *with* in old Catalan). Both husband and wife parts can be divided in two sub-parts, being his/her own information (e.g. name, surname, home-town) and the correspondent parents (*HPI* and *WPI*) information (*e.g.* father's name, deceased parents). These two last parts are connected by the keywords *fill de* and *filla de* (*son of* and *daughter of* in old Catalan) respectively. Formally, the syntactic rules are:

$$\begin{array}{ll} D \wedge H \wedge W & \Rightarrow L \\ DI \wedge rebere \wedge de & \Rightarrow D \\ HI \vee (fill \wedge de \wedge HPI) & \Rightarrow H \\ ab \wedge WI \vee (filla \wedge de \wedge WPI) & \Rightarrow W \end{array}$$

The semantic information is stored in the text word layout. The words are first categorized in several classes: *husband*, *wife*, *husband family*, *wife family* and other information. Each of these general class tags can be accompanied by more specific sub-class tags in a semantic and ontology-like way. In the XML example in Fig. 4 we show that *license 1* contains two classes: *husband* and *father husband*. In this case, each category contains a list of related tags: *husband name*, *husband surname*, *husband job* and *husband town*.

This class labelling system should not be seen as a complete parsing tree, since there are no word-level semantic labels (*e.g.* in a compound name, both words have the same label), but as a useful source of syntactic and semantic information. This semantic information can certainly be used to improve classic document analysis tasks approaches. But we believe that the real and novel value of this database is beyond this, as we will discuss next.

C. Uses of the database

Handwritten databases, either contemporary like *IAM* [17] and historical like *George Washington* [14], are generally designed for handwriting recognition. The main novelty of BH2M is that it covers the whole pipeline, including the specificity of the domain. So it also allows to test context dependent interpretation algorithms.

As stated in the introduction, this database has some distinctive folds and potentials. First of all, these highly structured documents require not only specific lexicons, but also specific language models. Secondly, and due to immigration, these documents contain a large amount of unknown new words (e.g. names, surnames, places), which are challenging for handwriting recognition approaches that deal with open vocabulary. Concretely, the test set contains a 32% of Out of Vocabulary Words (see Table I). Finally, these documents can be used for research in information extraction: the system can associate a semantic category to each word. Furthermore, they can represent the information contained in the marriage license using graphs in the form of keywords and relationships between them.

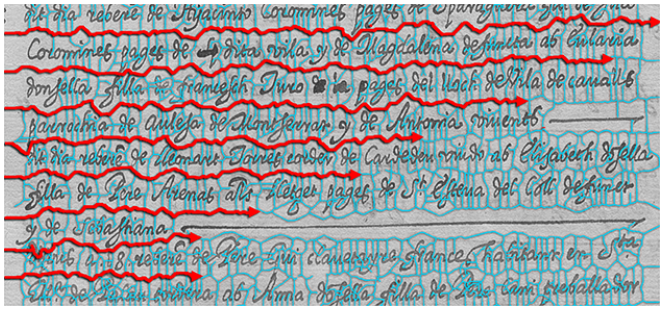

Fig. 5: Representation of the text line segmentation. The paths are located using the skeleton of the background.

In summary, in addition to the typical document analysis tasks, the presented database can be also used for research in knowledge extraction, cross-linkage and knowledge discovery.

III. Experimental Uses Cases

To illustrate the usefulness of the database to evaluate different tasks, we performed different experiments which can be seen as the most representative in this field. As an example of layout analysis, we show the results of a line segmentation approach. As an example of word recognition, we show the results of a segmentation-free and segmentation-based word spotting approaches.

The database has been divided into three parts for performance evaluation. We have randomly selected 100 pages for training, 34 pages for validation and 40 pages for testing. Table I contains the figures of the three defined partitions. The Out of Vocabulary Words (OOV) row shows the number of word classes in the test set that do not appear in the other partitions, as well as the number of word classes in the validation set that do not appear in the training set.

A. Line Segmentation

As baseline, we propose to use the line segmentation approach for Historical Handwritten Documents introduced [26].

This line segmentation approach is designed for dealing with the problems of touching components, curvilinear text lines and horizontally-overlapping components. The proposed algorithm formulates line segmentation as finding the central path in the area between two consecutive lines. This is solved as a graph traversal problem. A graph is constructed using the skeleton of the image background. Then, a path-finding algorithm is used to find the optimum path between text lines. In Fig. 5 we illustrate the algorithm.

One of the key purposes of this work is to solve the problem of touching lines adding new virtual edges to the graph. These characters are split using some heuristics which evaluate the spatial information around the area involved. This technique is not oriented to a specific writer, style or alphabet, and it is able to cope with multi-oriented text lines and historical documents.

Metric:	Test	All
N	1,266	5,368
M	1,295	5,557
one2one	1,053	4,388
Detection Rate (RT)	83.1%	81.7%
Recognition Accuracy (RA)	81.3%	78.9%
FM	82.1%	80.2%

TABLE II: Line segmentation results. The first experiment is using the *test* partition. The second one is using *all* the document pages.

1) Performance Evaluation: The evaluation of the line segmentation approach has been performed using the metrics from the ICDAR2013 Handwritten Segmentation Contest [27]. These metrics are based on counting the number of one-to-one matches between the areas that are detected by the algorithm and the areas according to the ground-truth. We will use a MatchScore table whose values are computed according to the intersection of the pixel sets of the result and the ground-truth.

Let G_i be the set of all points of the i ground-truth region, R_j the set of all points of the j result region, and $T(s)$ a function that counts the elements of set s. Table $MatchScore(i, j)$ represents the matching results of the i ground-truth region and the j result region as follows: $MatchScore(i, j) = T(G_i \cap R_j)/T(G_i \cup R_j)$. Detection Rate and Recognition Accuracy are calculated from the MatchScore table [28]. A region pair is considered as a one-to-one match only if the matching score is equal to or above the evaluator's acceptance threshold T_a. Let N be the count of ground-truth elements, M the count of result elements, and $o2o$ the number of one-to-one matches, the detection rate (DR) and recognition accuracy (RA) are defined as follows:

$$DR = \frac{o2o}{N}, \qquad RA = \frac{o2o}{M} \tag{1}$$

2) Results: We have performed the experiments using the approach described in Section III-A. The baseline method does not need any training process. Thus, in the first experiment we have only used the *test* set of the database. We have performed a second experiment using the three sets of the database (*train*, *validation* and *test* sets). The experimental results are shown in Table II. The results are quite similar in both experiments.

B. Word Spotting

We also propose a baseline for the evaluation or word spotting algorithm. In word spotting, the goal is to find all instances of a query word in a dataset of document images. One could classify the state-of-the-art approaches in two families: segmentation-based approaches [14], [29], which assume that the document image has been segmented into lines or words; and segmentation-free approaches [30], [10], which do not require any segmentation. For this reason, we present a baseline for each one of the representative approaches. The first approach is based on the traditional word spotting flow [14]. We represent the image words with a sequence of features [29], which result in a variable-length descriptor, and then, by using Dynamic Time Warping (DTW) as a similarity measure, word images in the dataset are compared to the query word and ranked according to this similarity. In this case we use the ground-truth information to segment the words from the document images. The second method [10] is based on a segmentation-free approach. It uses the exemplar-SVM framework, and relies on a sliding-window search to retrieve the document regions that are likely to contain the query word. In this case this method does not use the ground-truth information to crop the words in the documents.

1) Performance Evaluation: We use the following protocol to select the query words: each *non-stop* word[3] in the test set is considered as a query and used to rank all the regions of every document in the dataset. However, in order to reduce the bias in the results towards very frequent words, we limit the number of times that a word may be used as query. Concretely, if a word class has more than 50 occurrences, we randomly select 50 of these word images as query and we set them as the representative set for this word in all the experiments. This makes a total of 5,170 word image queries.

To evaluate the performance we combine the retrieved regions of all the documents and rerank them according to their score. The query image, if retrieved, is removed from the retrieved results and not considered in the performance evaluation. A region is classified as positive if its overlap over union with the annotated bounding box in the groundtruth is larger than 50%, and negative otherwise. We report the mean Average Precision (mAP) as our main measure of accuracy, which is a standard measure in retrieval systems and can be understood as the area below the precision-recall curve.

2) Results: We show the mAP obtained for segmentation-free and non-segmentation-free approaches in Table III. Note that in both cases only the set of queries proposed in Section III-B1 has been used to retrieve words from the test partition, and neither the train set nor the validation set have been used for this experiment.

Word Spotting	Approach	mAP
segmentation-based	DTW + Vinciarelli [29]	31.51
segmentation-free	HOG+EWS [10]	51.35

TABLE III: Word spotting results.

In addition to the query by example (QBE) methods presented as baseline, this dataset can also be used for query by string (QBS) searches. Although we do not show results of a QBS-based method as baseline, in order to ease the comparison between methods, we provide the set of text strings that should be used to query the database. In case of the QBS-based search, we use as queries all the unique *non-stop* text words that appears in the test set, *i.e.* each word is used only once as string query. This makes a total of 1,408 string words queries. We provide a list of both QBE and QBS queries along with the dataset.

IV. Conclusion

In this paper we have presented the Barcelona Historical Handwritten Marriages database. The data is compiled from

[3]We have carefully selected the list of *stop words* based on a commonly used list of Catalan stop words (http://latel.upf.edu/morgana/altres/pub/ca_stop.htm) and then adapted to the old Catalan.

a marriage license book collection and it is a useful tool for different research lines in handwriting document analysis. We have introduced a database that can be used for different research tasks, covering from layout analysis to text recognition. In addition to the classical document analysis tasks, the database can be used for research in knowledge extraction, cross-linkage and document understanding.

Along with the database, we provide baseline results for the main representative topics in handwritten document analysis: text line segmentation and word spotting. The baseline results demonstrate that these kind of documents are challenging. We hope that the presented ground-truth and baselines, can provide a benchmark for the research on historical document analysis in the near future.

As future work, we plan to increase the database with the index and other volumes of the same Marriage license books collection. We will introduce new volumes from different years and writers.

ACKNOWLEDGEMENT

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 and by two research grants of the UAB (471-01-8/09)

REFERENCES

[1] F. Fernandez and O. Terrades, "Document segmentation using relative location features," in *ICPR*, 2012, pp. 1562–1565.

[2] F. Álvaro, F. Cruz, J. Sánchez, O. Terrades, and J. Bened, "Page segmentation of structured documents using 2d stochastic context-free grammars," in *Pattern Recognition and Image Analysis*, 2013.

[3] M. Baechler, M. Liwicki, and R. Ingold, "Text line extraction using dmlp classifiers for historical manuscripts," in *ICDAR*, 2013.

[4] S. Bukhari, F. Shafait, and T. Breuel, "Towards generic text-line extraction," in *ICDAR*, 2013.

[5] K. Kumar and A. Namboodiri, "Learning segmentation of documents with complex scripts," *Indian conference on Computer Vision, Graphics and Image Processing*, 2006.

[6] A. Sarkar, A. Biswas, P. Bhowmick, and B. Bhattacharya, "Word segmentation and baseline detection in handwritten documents using isothetic covers," *ICFHR*, 2010.

[7] N. Cirera, A. Fornés, V. Frinken, and J. Lladós, "Hybrid grammar language model for handwritten historical documents recognition," in *Pattern Recognition and Image Analysis*, 2013.

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

[9] S. Marinai, "Text retrieval from early printed books," *IJDAR*, 2011.

[10] J. Almazán, A. Gordo, A. Fornés, and E. Valveny, "Efficient exemplar word spotting," in *BMVC*, 2012.

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

[12] T. Rath, R. Manmatha, and V. Lavrenko, "A search engine for historical manuscript images," *International conference on Research and development in information retrieval*, 2004.

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

[14] T. Rath and R. Manmatha, "Word spotting for historical documents," *IJDAR*, 2007.

[15] A. Fischer, A. Keller, V. Frinken, and H. Bunke, "Lexicon-free handwritten word spotting using character hmms," *PRL*, 2012.

[16] N. Serrano, F. Castro, and A. Juan, "The rodrigo database," in *LREC*, 2010.

[17] U. Marti and H. Bunke, "The iam-database: An english sentence database for off-line handwriting recognition," *IJDAR*, 2002.

[18] C. Liu, F. Yin, D. Wang, and Q. Wang, "Casia online and offline chinese handwriting databases," in *ICDAR*, 2011, pp. 37–41.

[19] M. Pechwitz, S. M. Snoussi, V. Mrgner, N. Ellouze, and H. Amiri, "Ifn/enit - database of handwritten arabic words," in *CIFED*, 2002.

[20] S. Athenikos, "Wikiphilosofia and pananthropon: Extraction and visualization of facts, relations, and networks for a digital humanities knowledge portal," in *ACM Conference Hypertext and Hypermedia (Hypertext 2009)*, 2009.

[21] D. J. Kennard, A. M. Kent, and W. A. Barrett, "Linking the past: discovering historical social networks from documents and linking to a genealogical database," in *HIP 2011*, 2011.

[22] D. Embley, S. Machado, T. Packer, J. Park, A. Zitzelberger, S. W. Liddle, N. Tate, and D. Lonsdale, "Enabling search for facts and implied facts in historical documents," in *HIP 2011*, 2011.

[23] T. Packer and D. Embley, "Cost effective ontology population with data from lists in ocred historical documents," in *HIP 2013*, 2013.

[24] M. Jindal, R. Sharma, and G. Lehal, "Segmentation of horizontally overlapping lines in printed indian scripts," in *IJCIR*, 2007.

[25] A. Amato, A. Sappa, A. Fornés, F. Lumbreras, and J. Lladós, "Divide and conquer: Atomizing and parallelizing a task in a mobile crowdsourcing platform," in *CrowdMM*, 2013.

[26] D. Fernández-Mota, J. Lladós, and A. Fornés, "A graph-based approach for segmenting touching lines in historical handwritten documents," *International Journal on Document Analysis and Recognition (IJDAR)*, pp. 1–20, 2014.

[27] N. Stamatopoulos, B. Gatos, G. Louloudis, U. Pal, and A. Alaei, "Icdar 2013 handwriting segmentation contest," in *ICDAR*, 2013.

[28] I. Phillips and A. Chhabra, "Empirical performance evaluation of graphics recognition systems," *Pattern Anal. Mach. Intell.*, 1999.

[29] A. Vinciarelli and S. Bengio, "Offline cursive word recognition using continuous density hidden markov models trained with pca or ica features," in *ICPR*, 2002.

[30] M. Rusiñol, D. Aldavert, R. Toledo, and J. Lladós, "Browsing heterogeneous document collections by a segmentation-free word spotting method," in *ICDAR*, 2011.