SHIBR-The Swedish Historical Birth Records: A Semi-Annotated Dataset

Abbas Cheddad[1*], Hüseyin Kusetogullari[1‡], Agrin Hilmkil[2‡], Lena Sundin[3], Amir Yavariabdi[4], Mustapha Aouache[5], Johan Hall[6]

[1] Department of Computer Science, Blekinge Institute of Technology, SE-371 79, Karlskrona, Sweden
[2] Peltarion AB, Holländargatan 17, SE-111 60, Stockholm, Sweden
[3] Independent Researcher, Stockholm, Sweden
[4] Department of Mechatronics Engineering, KTO Karatay University, Konya, Turkey
[5] Division Télécom, Centre de Développement des Technologies Avancées (CDTA), BP 17, Baba-Hassen, Algiers, Algeria
[6] Arkiv Digital AD AB, Växjö, Sweden
[*] Corresponding author
[‡] These co-authors contributed equally to the manuscript

Abstract This paper presents a digital image dataset of historical handwritten birth records stored in the archives of several parishes across Sweden, together with the corresponding metadata that supports the evaluation of document analysis algorithms' performance. The dataset is called SHIBR (the Swedish Historical Birth Records). The contribution of this paper is twofold. First, we believe it is the first and the largest Swedish dataset of its kind provided as open access (15,000 high-resolution colour images of the era between 1800 and 1840). We also perform some data mining of the dataset to uncover some statistics and facts that might be of interest and use to genealogists. Second, we provide a comprehensive survey of contemporary datasets in the field that are open to the public along with a compact review of word spotting techniques. The word transcription file contains 17 columns of information pertaining to each image (e.g., child's first name, birth date, date of baptism, father's first/last name, mother's first/last name, death records, town, job title of the father/mother, etc.). Moreover, we evaluate some deep learning models, pre-trained on two other renowned datasets, for word spotting in SHIBR. However, our dataset proved challenging due to the unique handwriting style. Therefore, the dataset could also be used for competitions dedicated to a large set of document analysis problems, including word spotting.

Keywords: Historical data of birth records, handwritten documents, public dataset, word spotting

1 Introduction

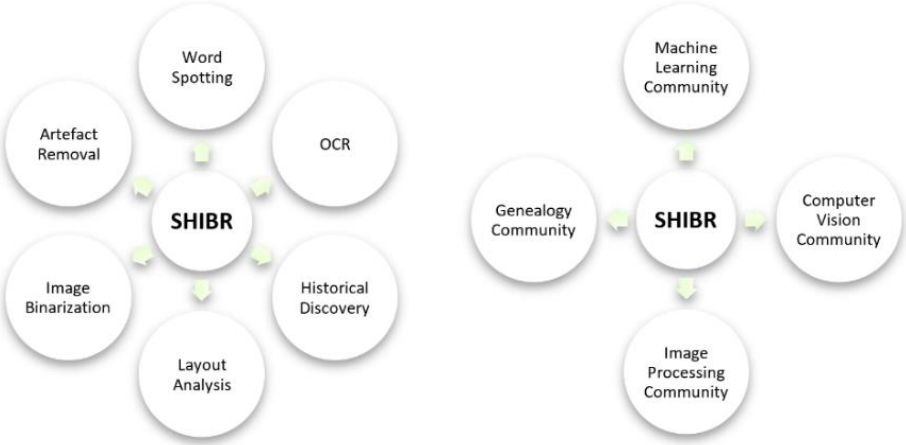
Digitising the past is a way to preserve history, restore deteriorating/uncompleted text, extract facts and information, and help in searching, document retrieval and data mining tasks. The digitisation of books/documents is among the objectives that current digital libraries and electronic government initiatives are putting on the top of their priorities. For example, dozens of universities, research centres and companies in Europe have together started a large-scale EU consortium called IMPACT[1] (Improving Access to Text) [1][2]. Among the "*Endangered Archives Programme*" initiatives of the British Library is the digitisation of manuscripts of the Al-Aqsa Mosque Library, East Jerusalem [3]. This historical collection contains more than a hundred Arabic language titles that span over several Islamic periods from the 9th century CE to the end of the Ottoman rule in Palestine at the beginning of the 20th century. These books span topics about the Arabic language and literature, logic, math, religion, and Sufism[2].

An old and still valid way to transcribe historical handwritten documents is to rely on crowdsourcing. It is the practice of gathering information or input into a task by acquiring the services of a large number of people (a.k.a. crowd). It is often characterised by small and short-term deals [4]. In a recent study, crowdsourcing, when

[1] IMPACT: http://www.impact-project.eu/home/, [Online], accessed on 2020-04-11.
[2] https://eap.bl.uk/collection/EAP521-1, [Online], accessed on 2020-04-11.

combined with contemporary technology, is shown to deliver far more complete and validated data than automated processes alone could produce [5]. As such, the automated process of converting a historical document into a readable text is still posing various challenges. The work herein offers possible assistance in lifting some of these challenges by providing one of the most extensive free-access semi-annotated historical handwritten document datasets.

We conclude this section by noting that this dataset would enrich the availability of historical handwritten document datasets and help develop more accurate algorithms for word spotting, optical character recognition (OCR), document layout analysis and image binarization. It would also serve the research community interested in history and heritage (i.e., genealogists), see Fig.1. This set of motifs was the ultimate impetus for the creation of the SHIBR dataset (the **S**wedish **Hi**storical **B**irth **R**ecords)[3]. This complete-page dataset (SHIBR) complements the previously published numerical handwritten dataset (ARDIS) [6], both of which are generously provided for free by *Arkiv Digital AD AB*, a Swedish company.



(a) (b)

Fig. 1 SHIBR and what it serves: per discipline (a) and per community (b)

2 Review of Related Public Datasets

Different public handwritten document image datasets have been created and presented to resolve various document image challenges such as text line segmentation [7], word spotting [8], writer identification [9], digit and character segmentation and recognition [10 - 12], binarization [13], and a variety of other challenges [14 - 16]. These datasets enable researchers to develop automated and computationally efficient algorithms. Generally, the existing datasets are classified into two groups based on the era in which they were written: historical or modern datasets. The well-known and widely used document databases are listed in Table 1, several of which are described in this section.

George Washington database (GW) [17, 18]: This database is a baseline database for text line segmentation, word spotting and word recognition tasks. The Washington database consists of 20 historical handwritten document images written in English with longhand script and ink-type pen in the 18th century. Moreover, these images are annotated with 656 text lines, 4894-word instances, 1471-word classes and 82 letters.

IAM database [22, 23]: The IAM database contains 1539 handwritten modern document images written by 657 different English scriptwriters. The documents were scanned at a resolution of 300dpi and stored in greyscale colour to create this database. The document images are labelled using an automatic segmentation approach and

[3] The SHIBR data set is open access available from the following link: https://ardisdataset.github.io/SHIBR/

are verified visually. The database consists of 5685 isolated and labelled sentences, 13353 isolated and labelled text lines, 115320 isolated and labelled words.

Table 1 Annotated handwritten image datasets in different languages that are publicly available

Dataset name	**Language**	**Context**	**Total images**	**Colour Type**
George Washington [17, 18]	English	Historical	20	RGB
CMATERdb1 [19]	Bangla–English	Historical	150	RGB
H-KWS-2016 Bentham [20, 21]	English, German	Historical	279	RGB
IAM [22, 23]	English	Modern	1539	Greyscale
VML-HD [24]	Arabic	Historical	668	RGB
HADARA80P [25]	Arabic	Historical	80	RGB
BADAM [26]	Arabic	Historical	400	RGB
Esposalles [27, 28]	Spanish	Historical	173	RGB
IFN/ENIT [29]	Arabic	Historical	6735	RGB

VML-HD database [24]: This database includes 668 handwritten document images. The documents in the database were written in Arabic by different writers between the years 1088 and 1451. All the words and characters in the document images are manually annotated with bounding boxes. As a result, this database consists of 159149 annotated words and 326289 annotated characters.

HADARA80P database [25]: This database contains 80 handwritten document images written in the Arabic language. This database is used for word segmentation problems, as the words in the document images are annotated with polygons.

Esposalles database [27, 28]: The Esposalles database is a Spanish historical handwriting document image database consisting of 173 document images. The documents were written between 1451 and 1905, and they contain information from the marriage licenses of Spanish citizens. These documents are collected from different books available at the archives of the Cathedral of Barcelona. All text blocks, lines, and transcriptions in the document images are manually labelled. Furthermore, this database has been used to develop handwriting recognition algorithms.

Other handwritten document image databases have also been created and can be found with more details in [19, 21, 26, 29]. Furthermore, the proposed SHIBR document dataset is comprehensively scrutinised and described in section 4.

2.1. Limitations of existing document images databases

Even though some of the existing datasets have annotations, most of them have several limitations: 1) scarcity of a large number of document images; 2) lack of datasets with Swedish characters; 3) lack of availability of historical documents written in Swedish handwriting styles with various types of dip pens; and 4) lack of availability of datasets with significant variations of artefacts (e.g., degradation, bleed-through, ink leakage etc.). For instance, the George Washington dataset contains the least number of document datasets with 20 English document images. In contrast, the IFN/ENIT dataset, the most extensive document image database (see Table 1), consists of 6735 binary Arabic document images. Therefore, the main challenges when using these datasets for historical document image analysis are 1) dealing with a small number of document images which leads to small intra- and inter-class variations, 2) exhibiting a small number of artefacts, 3) covering historical ancient handwriting styles insufficiently. Therefore, to support the development of research in historical document image analysis, it is essential to construct a new dataset that would address the shortcomings of the existing datasets. Thus, this paper proposes a new and large dataset (SHIBR) containing 15000 document images, which is the largest of its kind as far as we are aware. The SHIBR dataset is semi-annotated, easing the development of automated and semi-automated machine learning methods for document analysis applications. Furthermore, to the best of our knowledge, the SHIBR dataset is the largest historical handwritten document dataset and the first semi-annotated historical document image dataset with Swedish characters.

3 Challenges and Opportunities in Historical Handwritten Documents

In the context of handwritten document image analysis, many challenges need to be tackled [30]. Most of the state-of-the-art solutions focus mainly on word spotting and recognition challenges.

3.1. Word/Pattern Spotting

Over the past decade, an enormous collection of handwritten or machine-printed documents have been digitised to preserve the contained information. Word spotting methods are used to extract relevant information from these documents. Generally, word spotting in handwritten document images is much more complex than on machine-printed document images—the former consists of significant variations of handwriting styles and various character types in different languages. As the text in these historical documents was written by different writers, it generates large variations in appearances because of, on the one hand, skewness, curvature, aspect ratio, and size, and because of broken and connected words/characters on the other hand. As a result, these variations in writing styles in different languages may create endless diversities for word spotting in handwritten document images. Many word spotting methods have been proposed for document indexation. They can be classified into two groups: 1) segmentation-based and 2) segmentation-free methods.

Matching is a segmentation-free word spotting approach. It is the process of searching a target or template word image in document images. It is one of the basic approaches that have been applied for word spotting on document images. Moreover, it is mainly employed based on similarity or distance measure between the template word image cropped from a document image and the document images' target region. In [31], a word-level matching scheme is proposed to search a template word image in printed document images using a feature-extraction technique. After that, the extracted features are used for similarity estimation for word spotting. In [32], another word spotting framework is proposed; the dynamic time warping (DTW) based matching technique. The DTW matching algorithm is applied on machine-printed document images, providing superior results for word spotting. In [33], a block adjacency graph (BAG) method for word spotting is designed and employed based on similarity estimation between the template image and the moving window regions in document images. A word shape coding scheme is proposed by Bai et al. [34] that combines feature descriptors and a matching technique for word spotting in document images. A block-based document image descriptor used for word spotting in historical printed documents based on the template matching process is proposed by Rabaev et al. [35]. Their experiments show that this method provides promising results if the documents do not include too many undesired artefacts. Other word spotting based matching methods can be found in [36 - 38]. The matching based word spotting techniques have several drawbacks: 1) they are time-consuming, 2) they cannot overcome undesired artefacts involved in the handwritten documents, and 3) they often yield poor accuracy rates on handwritten text images.

Thus, learning-based segmentation-free word spotting techniques are designed and applied to increase word spotting accuracy. For instance, the Hidden Markov Models (HMMs) technique has been used for word spotting in handwritten documents [39, 40]. Besides, hybrid models of HMMs with different supervised learning methods have been developed that combine HMMs with Support Vector Machine (SVM) [41] or with Neural Network (NN) [42] or with deep Convolutional Neural Network (CNN) [43]. In another work, a new word spotting system for handwritten Urdu language document images is proposed [44]. The method uses several pre-processing steps such as binarization, connected component analysis and edge detection. Subsequently, for word spotting purposes, a sliding window based on an SVM classifier is used to spot Urdu words. In [45], a word spotting and recognition approach based on a common representation of word images and text strings is proposed. The method first extracts the standard features to decrease the dimensional space, and then the nearest neighbour algorithm is used for word spotting. Frinken et al. [46] have designed a novel method based on recurrent neural network (RNN) for word spotting in handwritten documents. In [47], an efficient patch-based framework combined with the scale-invariant feature transform (SIFT) descriptors is proposed for keyword spotting in historical document collections. In [48], a CNN architecture is designed for word spotting in handwritten documents. Extensive survey papers of words spotting methods can be found in [49 - 53].

SHIBR lends itself well as a challenging benchmark for word spotting methods. Since the SHIBR dataset, like many other historical documents, does not have segmented words, we exclusively look at the segmentation-free

methods. These methods may, for example, rely on traditional computer vision to find interesting regions or on region-proposal networks like Faster-RCNN [54]. The Ctrl-F-Mini algorithm fulfils this criterion and has been shown to outperform many existing methods [55]. We, therefore, select Ctrl-F-Mini for the first benchmarks on SHIBR. While its dependence on bounding boxes during training makes it difficult to train on SHIBR, its pre-trained models may be readily evaluated on the dataset. Ctrl-F-Mini is a deep convolutional network model for segmentation-free word spotting. It is related to the Faster-RCNN alternative but uses Dilated Text Proposals (DTP) [55] instead of a Region Proposal Network (RPN) [54] to propose regions of a manuscript page potentially containing words. For each region, the network outputs the estimated probability of it depicting a word and a word embedding. The word embeddings are either the Pyramidal Histogram of Characters (PHOC) or the Discrete Cosine Transform of Words (DCToW) [55] embeddings. In word retrieval, the regions are ranked by their embeddings' cosine similarity to the query string embedding.

3.2. Opportunities: Genealogy Research

Genealogy and the study of family history/tree are both interlinked. In the modern era, and with the success and spread of DNA sequencing in high-throughput genomic sequences, genealogy has become a vivid field. However, genealogists are also interested in unravelling the history of families by mining historical documents. Nevertheless, there exist contentious points that Hatton [56] has rethought in a study examining history, lineage, identity, and technology in relation to genealogy. As depicted in Fig. 2, the top twenty countries most interested in genealogy show diversity, with Sweden coming slightly above the USA. The statistics are retrieved from the *Google Trends* tool (a website by Google that analyses the popularity of top search queries in Google Search across various regions and languages).

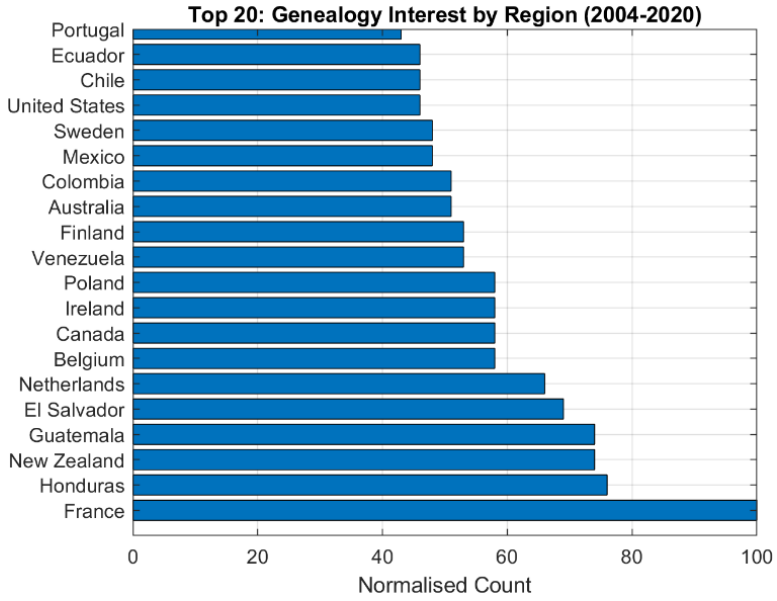


Fig. 2 Top twenty countries most interested in genealogy as recorded by the *Google Trends* tool between 2004-2020

3.3. Opportunities: Window into the Past

A window into the past may allow the investigation of assumptions about the historical position and actual practices and thinking of a particular epoch, presenting genealogists with opportunities to further their theories. For instance, Abildgren, K. explores what a crowdsourced genealogical online database can say about Denmark's income inequality during the First World War [57]. Additionally, Zhu Z. constructs the concept system and relationships of the genealogy ontology and takes Wu Shilai, the ancestor of the 23rd generation of Wu's, as an example to realize the visualisation of traditional Chinese genealogy [58].

4 SHIBR Dataset

This dataset is retrieved from the *Arkiv Digital AD AB* image and index database. When a child was born in Sweden in the 1800s, he or she was registered by a priest in a church record book called *Birth and Christening Records*. These priests registered the child's name, when the child was born and baptised, where the child lived, and information about the father and the child's mother, as shown in Table 2. The transcription is based on manual annotation (at *Arkiv Digital AD AB* and its partners) of scanned images from 1800 to 1840.

4.1. Structure of SHIBR

The master dataset ($SHIBR_m$) consists of 818,110 indexed rows and 64,084 images. This dataset is confidential and can only be used according to the agreement between the company *Arkiv Digital AD AB* and the *Blekinge Tekniska Högskola* (BTH). However, a subset comprising random samples from the period 1800-1840 was determined to abide by the GDPR (the European General Data Protection Regulation) law and, therefore, can be made open access. As such, the public dataset ($SHIBR_p$) with semi-annotation consists of 10500, 2250, 2250 images for training, testing and validation, respectively. Hence, in total, the $SHIBR_p$ public dataset consists of 15,000 high resolution (2000x1300 to 6000x4000) images in RGB (Red, Green and Blue) colour space (~50GB of data) that exhibit a variety of layouts, handwriting styles, background colour and degradations. Additionally, $SHIBR_p$ is associated with *Excel* spreadsheets for each of the three folders (training, testing and validation). In total, the spreadsheets contain 191,301 entries.

Swedish counties (län) covered: The counties that are recorded in these books are as follows. Gotland, Gävleborg, Norrbotten, Västerbotten, Västernorrland, Västmanland, Älvsborg[4], and Örebro (see Fig. 3).

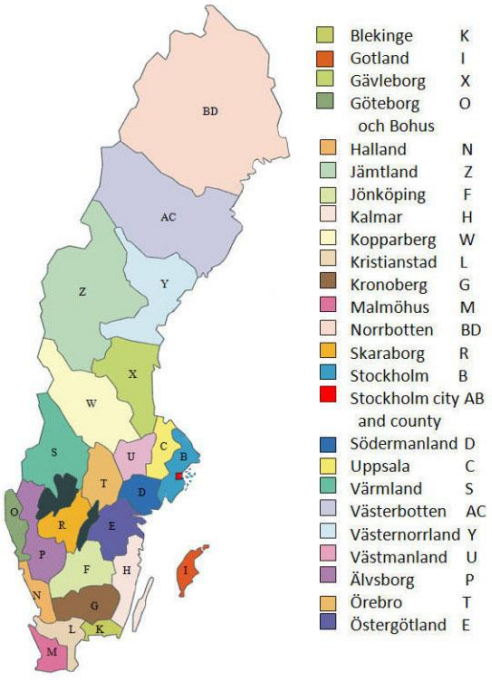


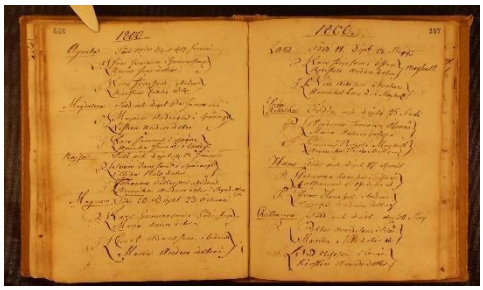
Fig. 3 Map of Sweden, as before the year 1997 [Source: Familysearch.org], showing some of the Swedish counties reported in the SHIBR dataset

[4] Älvsborg county was a county of Sweden until 1997 when it was merged with the counties of Göteborg, Bohus and Skaraborg to form Västra Götaland County. The county consisted of the provinces of Dalsland and the central part of Västergötland with the seat of residence in the city of Vänersborg.

Description of the index columns: Each image (of the 15,000) corresponds to a double page of a book, and each of these images is associated with an entry in the annotation file (manual transcription) with 17 columns, as shown in Table 2. The structure of the transcribed file, along with a sample image, is exemplified in Fig. 4.

Table 2 Columns in $SHIBR_p$'s accompanying transcribed file

Column	Column Name	Description
1	id	Company's ID in the database
2	index_aid	Index AID (Company's external ID)
3	county	County where the child was born or registered (usually not in the image)
4	parish	Where the child was born or registered (can be written at the top of the page or entirely missing from the image)
5	child_first_name	Given name of the child
6	birth_date	Date of birth, format YYYYMMDD (usually written DD/MM with the year on top of the page)
7	baptism_date	Date of baptism, format YYYYMMDD (usually written DD/MM with the year on top of the page)
8	birth_place	Place of birth
9	father_title	Title or occupation of the father
10	father_first_name	Given name of the father
11	father_last_name	Surname of the father
12	mother_title	Title or occupation of the mother
13	mother_first_name	Given name of the mother
14	mother_last_name	Surname of the mother
15	mother_age	Age of the mother when the child was born
16	image_aid	Image AID (Company's external image ID)
17	image_path	Relative path to the image (images/<image_path>)



id	index_aid	county	parish	child_first_name	birth_date	baptism_date	birth_place	father_title	father_first_name	father_last_name	mother_title	mother_first_name	mother_last_name	mother_age	image_aid	image_path
172073263	r11.p172073263	Örebro	Almby	Anna Stina	18000115	18000116	Örebro Slåttet				pigan	Lisa	Olofsdotter		v52844.b59.	52844/v52844.b59.s109.jpg
172073264	r11.p172073264	Örebro	Almby	Malena	18000131	18000204	Örebro	hemmansbri	Lars	Larsson	husfru	Stina	Pehrsdotter		v52844.b59.	52844/v52844.b59.s109.jpg
172073392	r11.p172073392	Örebro	Almby	Olof	18061001	18061005	Sörby	rusthållaren	Jan	Ersson	husfru	Maja Stina	Zachrisdotter	27	v52844.b72.	52844/v52844.b72.s135.jpg
172073393	r11.p172073393	Örebro	Almby	dödfödd son	18061001		Almby		Anders	Andersson	husfru	Anna Greta	Andersdotter		v52844.b72.	52844/v52844.b72.s135.jpg
172073394	r11.p172073394	Örebro	Almby	Johan	18061026	18061027	L. Gräflinge	torparen	Johan	Nilsson	husfru	Stina	Amundsdotter	42	v52844.b72.	52844/v52844.b72.s135.jpg
172073395	r11.p172073395	Örebro	Almby	Anna Catharina	18061028	18061102	Ormesta Ägc	sockne smed	Fredric	Ringström	husfru	Maja Lisa	Olsdotter	28	v52844.b72.	52844/v52844.b72.s135.jpg
172073396	r11.p172073396	Örebro	Almby	Lena Christina	18061202	18061204	Tyble	hälftenbruk.	Niclas	Jacobsson	husfru	Anna Cajsa	Andersdot.	21	v52844.b72.	52844/v52844.b72.s135.jpg
172073397	r11.p172073397	Örebro	Almby	Eric	18061202	18061203	St. Nybygget	torparen	Eric	Ersson	husfru	Maria	Ersdotter	43	v52844.b72.	52844/v52844.b72.s135.jpg
172073398	r11.p172073398	Örebro	Almby	Olof	18061215	18061216	Näsby Ägor	inhyses man	Anders	Olsson	husfru	Anna Greta	Carlsdotter	40	v52844.b72.	52844/v52844.b72.s135.jpg

Fig. 4 Example of a scanned book page and the view of a file containing transcribed data (17 columns). See Table 2 above for descriptions of columns 1-17

4.2. Mining $SHIBR_m$ – Statistical insights

Data mining methods aim at discovering frequently occurring patterns in a source dataset [59]. Here, we deploy simple basic statistics to identify potentially helpful information associated with the document images in the $SHIBR_m$ pertaining to the 19th century's era. The findings listed here are merely examples of the uncharted side of the $SHIBR_m$ dataset and of what its public version, $SHIBR_p$, can offer to the research community, especially to genealogists. Please note that the statistics drawn herein are only reflections of the set of data we currently have (i.e., $SHIBR_m$). In no way should they be taken as a de facto reference to the total population's overall reality during that era. We only analyse and describe the data that we have.

- *Birth rate stratified by county/year*: By examining the birth rate, we can see that it has an overall increasing trend and seemingly aligned with the public statistics[5]. When the retrieved data is stratified

[5] Sveriges folkmängd från 1749 och fram till idag [Sweden's population from 1749 until today] by the National Central Bureau of Statistics (SCB), 2017-10-27. Available from: https://www.scb.se/

by the county, we see, at the macro-level, that Älvsborgs län exhibits a large birth rate, as shown in Fig. 5. For example, this could be linked to socioeconomic status.

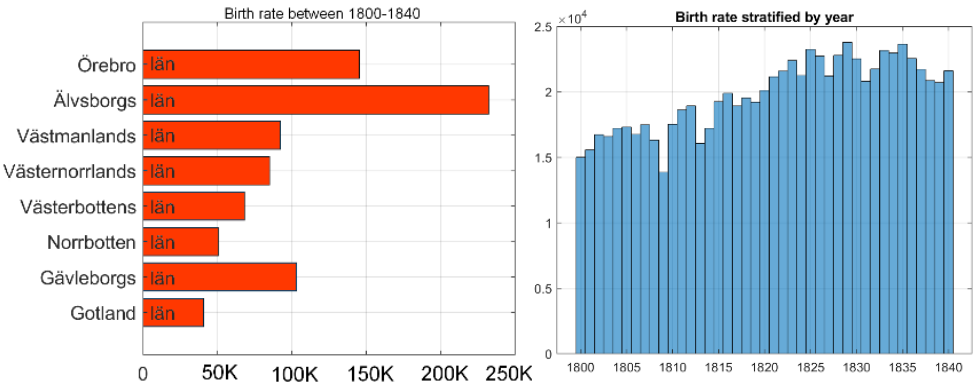


Fig. 5 Birth rate aggregated from the eight counties (right) and stratified by county (left)

- *Rate of stillborn (dödfödd)*: Analogues to the birth rate, and probably driven by it (dependent variables), is the death rate which also exhibits an increasing trend at the macro-level. Table 3 shows that Gotlands län tops the list with 1.857% of total newborns, and at the bottom of the list is Norrbottens län with 0.749%. Worth noting is that baby boys consistently exhibit higher death rates than baby girls in the data we have spanning the period 1800 to 1840. We are uncertain if this difference is genuine and descriptive of the total population in that period. However, this finding is consistent with a recent report published by the Statistics Sweden SCB (a Swedish agency) stating that "*infant mortality has been higher for boys than for girls but this difference between the sexes is almost non-existent in the 21st century*" and with the Historical Statistics of Sweden [60]. Fig. 6 shows the overall mortality rate stratified by gender, and Table 3 tabulates the rate at each county.

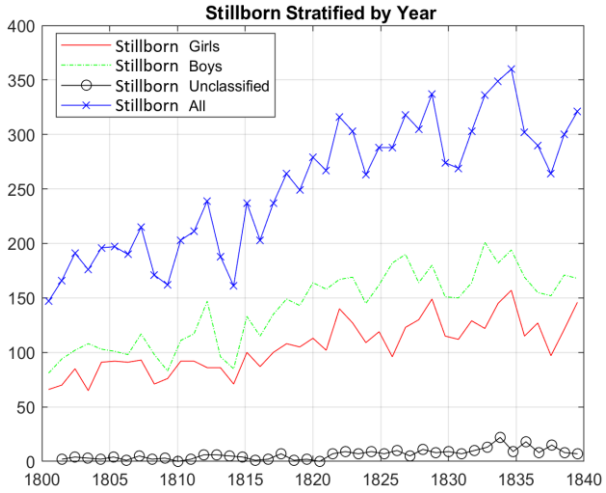


Fig. 6 Stillborn rate in all counties

Table 3 Stillborn rate stratified by counties

Län (county)	**Born Babies**	**Dead**	**%**
Gotlands län	40870	759	**1.857**
Gävleborg län	103364	1677	1.622
Norrbottens län	50608	379	**0.749**
Västerbottens län	68229	626	0.917
Västernorrlands län	85000	1131	1.331
Västmanlands län	92456	1589	1.719
Älvsborgs län	232101	2598	1.119
Örebro län	145482	1876	1.290

- *Period until baptised*: Baptism is a Christian rite for acceptance and adoption into Christianity. It was socially unacceptable not to baptise a child during that era. $SHIBR_m$ stores the number of days from birth to baptism. Fig. 7 illustrates Älvsborgs län as the county with a minor time interval between birth and baptism. Norrbottens län and Västerbottens län top the list, probably because being located in northern Sweden, where a large part of the population is made up of the Sámi people (Laplanders in English) who may have likely had to travel long distances to churches for baptism.

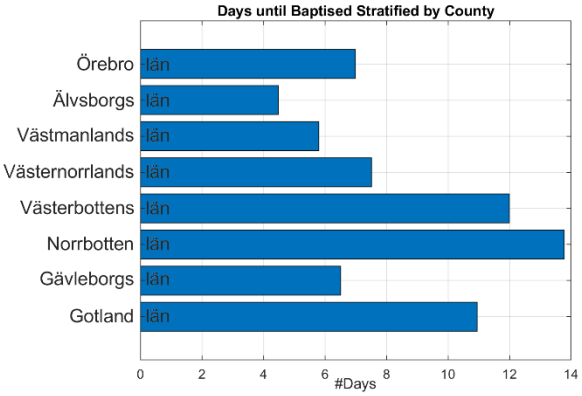


Fig. 7 Days between a baby is born until being baptised, stratified by counties

- *Most common first names (babies, women, men)*: Another typical trend to look at is the popularity of the first names given to babies and parents. Some of these names are fading away in modern Sweden, allowing for more trendy ones, especially among the young generation. Table 4 shows the top 10 most common names among newborns, fathers and mothers.

Table 4 Ten most common names in Sweden in the period between 1800 and 1840

Baby Girls		Baby Boys		Mothers		Fathers	
Anna	81645	Anders	51987	Anna	141541	Anders	86995
Brita	28659	Carl	32720	Brita	69245	Olof	62462
Maria	27172	Johan	30706	Stina	66090	Lars	57416
Johanna	22893	Lars	28050	Maria	48998	Eric	49772
Christina	18671	Johannes	25035	Maja	32366	Pehr	39945
Stina	17688	Olof	24265	Greta	27168	Nils	36629
Maja	11367	Eric	23428	Cajsa	22930	Jan	30283
Sara	10598	Pehr	19613	Sara	21293	Johan	29216
Catharina	9739	Jonas	18029	Lisa	20538	Jonas	21639
Carolina	8311	Nils	15838	Catharina	19116	Carl	21390

- *Age of women in the birth records*: What is the average age of mothers in the birth records of $SHIBR_m$ during the period 1800-1840? Fig. 8 depicts a bar chart of the age distribution of women in all counties. More in-depth statistics stratified by counties are tabulated in Table 5.

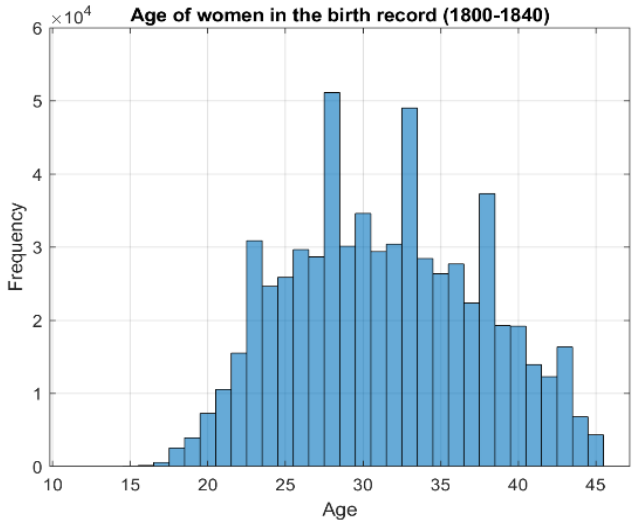


Fig. 8 Overall age distribution of women in the birth records (1800-1840)

Table 5 Age statistics of women in the birth records stratified by counties (1800-1840)

County (Län)	Mean	Standard deviation	Median	Mode
Gotlands län	30.751	6.177	30	28
Gävleborgs län	31.585	6.023	31	29
Norrbottens län	31.665	6.228	32	28
Västerbottens län	31.322	6.150	31	28
Västernorrlands län	31.699	6.035	32	28
Västmanlands län	31.174	6.183	31	28
Älvsborgs län	31.614	6.236	31	30
Örebro län	31.415	6.266	31	28

- *Most typical job titles (women/men)*: A final factor we look at in $SHIBR_m$ data is the job title of men and women during that period. As can be seen from Table 6, most of the men were farmers, military officers, government employees (tax officers), etc. Women during that period were not empowered to have their own jobs, and thus we see a lack of job titles in their column. Maids and housekeepers were the only reported jobs we found in the large dataset $SHIBR_m$.

Table 6 The most common job titles for men/women during the period (1800-1840)

Men		Women	
Swedish	**English**	**Swedish**	**English**
Dr./Dräng	Farm labourer	Pigan	Maid, female equivalent of dräng with slight variation in duties
Gårdssmeden	Smith	Husfru	Female head of household
Befallningsmannen	Head of region or farm	-	
Skatt/Skatten	Tax office worker	-	
Gjutar/Gjutaren	Casting worker	-	
Hemmansägaren	Yeoman	-	
Hammarsmedsmästare	Trip hammer worker	-	
Husar/Husaren	Cavalry soldier	-	

Back in the days, the Swedish written language was not standardized like it is today. Most people could not read and write. In 1842 there was a law enforcing that every child must go to primary school in Sweden. Only certain groups of people could read and write, for example, priests. As such, the spelling of words and names could be different across Sweden and between individuals. This asymmetry exhibits heterogeneity in both spelling and writing. Historical texts heterogeneity adds more complexity to perform robust image processing and recognition tasks. However, after the first-round manual transcription of the SHIBR dataset by the company's (Arkiv Digital AB) partners, the company carried out a validity check by Swedish native speakers to improve the transcription quality.

5 Experiments and Results

In this section, we present experimental results of the chosen algorithm's performance, Ctrl-F-Mini [55] (discussed in section 3.1), on page retrieval in SHIBR. In all cases, we use the trained models and implementation made available by the original authors. These models were trained on the ADAM dataset with a learning rate of 0.001, multiplied by 0.1 at every 10,000 steps. After a total of 25,000 steps, the model with the highest hold-out-set score was used. While the Ctrl-F-Mini comes with models trained with different loss functions, we select the ones trained with the cosine loss as they have the best benchmark results. However, we evaluate the PHOC and the DCToW embedding variants. Finally, since one model is provided per fold, we always select the one corresponding to the first fold.

The $SHIBR_p$ scanned pages are converted to greyscale colour space and pre-processed using the model's provided settings for the George Washington dataset. We use an NMS (Non-Maximum Suppression) threshold of 0.4 for the DTP regions for all experiments as this corresponds to the negative-match threshold in Ctrl-F-Mini [55]. We do not use any additional word-likeness filtering, and we do not limit the number of regions per document before predicting the embeddings.

5.1. Segmentation-free evaluation of word spotting

The most used metric for evaluating word spotting is the Mean Average Precision (mAP) [61]. For a given query q, the precision at k is:

$$P_k(q) = \frac{|\{top\ k\ candidates\ for\ q\} \cap \{instances\ of\ q\}|}{k} = \frac{1}{k}\sum_{j=1}^{k} r_j(q), \quad (1)$$

where the relevance indicator $r_j(q)$ is 1 if the candidate with rank j matches q and 0 otherwise. Averaging the value over all possible k gives the Average Precision (AP)

$$AP(q) = \frac{\sum_{k=1}^{n} P_k(q) r_k(q)}{|\{instances\ of\ q\}|}, \quad (2)$$

To finally retrieve the mAP, the AP is averaged over the set of all queries Q for the task:

$$mAP(Q) = \frac{1}{|Q|}\sum_{q \in Q} AP(q), \quad (3)$$

The mAP is calculated with instances of individual words to evaluate word spotting on datasets with bounding boxes. The relevance indicator $r_k(q)$ is determined by the word matching to ensure a particular overlap between the retrieved candidate and the ground truth bounding box. Since $SHIBR_p$ does not contain bounding boxes, this is not feasible.

To evaluate a word spotting algorithm on semi-annotated datasets like $SHIBR_p$, we instead propose evaluating the mAP with respect to page retrieval (mAP_{page}). Instances are therefore scanned pages instead of individual words. The set of queries is chosen to be the set of all words occurring in the text. All queries are performed over all the pages, and the results are then ranked according to their single best match for a given query. Finally, relevance $r_j(q)$ is indicated if q occurs on a page at rank j.

5.2. Results

As a baseline on the $SHIBR_p$ test set, we compute the mAP_{page} with respect to page retrieval (section 5.1) using the Ctrl-F-Mini models trained on the George Washington Dataset [17] and on the IAM Offline Handwriting Dataset [23]. We use the 100 most commonly occurring words in the test set as the set of queries Q. The results, including a random baseline corresponding to randomly ranking all instances for each query, are presented in Table 9.

Table 9 mAP_{page} with respect to page retrieval on the $SHIBR_p$ test set described in Sections 5.1 and 5.2. The result for the random baseline is the mean ± standard deviation over 100 trials

Model	**Training dataset**	**mAP_{page} on $SHIBR_p$**
Ctrl-F-Mini with PHOC	George Washington [17]	22.0%
Ctrl-F-Mini with DCToW	George Washington [17]	22.7%
Ctrl-F-Mini with PHOC	IAM [23]	22.8%
Ctrl-F-Mini with DCToW	IAM [23]	22.8%
Random baseline	-	(21.4 ± 0.1) %

As might be expected, the models trained on the much larger IAM dataset are consistently slightly better than those trained on the George Washington dataset. However, none of the results significantly outperform the random baseline. We would like to highlight two particular challenges that could contribute to the poor results. First, we acknowledge that none of the queries, Swedish names, are present in the training sets. Previous results have indicated that out-of-vocabulary queries tend to be more difficult for word spotting models [55]. Second, none of the provided models has been exposed to Swedish characters or accounted for in the embeddings. Third, the cursive and connected nature of text lines in our dataset and the variation in the writing style hinder achieving robust results using existing pre-trained models [62]. Building generally applicable word spotting applications requires overcoming these challenges. Furthermore, since real-world applications (e.g., when dealing with big data) also perform page retrieval, the mAP_{page} represents an essential and complementary metric for evaluating word spotting methods.

6 Conclusion

This paper contributes to the research community by providing an open-access large dataset of historical handwritten documents. These documents result from a continuous effort to digitise church birth books available from parishes across Sweden. The dataset that we provide comprises 15,000 high-resolution images in RGB colour space and transcribed files containing a wealth of information spanning a period of time from 1800 to 1840. The main goal of sharing the SHIBR dataset with the research community is to spark more initiatives to further develop robust document analysis algorithms (e.g., word spotting, document retrieval, character recognition, binarization, layout analysis, etc.) and to promote cross-disciplinary research studies.

Acknowledgements

This project is funded by the research project "*DocPRESERV: Preserving & Processing Historical Document Images with Artificial Intelligence*", STINT, the Swedish Foundation for International Cooperation in Research and Higher Education (Grant: AF2020-8892) and by the research project "*Scalable Resource Efficient Systems for Big Data Analytics*", the Knowledge Foundation (Grant: 20140032) in Sweden.

We also acknowledge the support of the Swedish company, *Arkiv Digital AD AB*, for providing the SHIBR dataset and for allowing us to make it open access.

Finally, we acknowledge the editorial committee's support and the insightful comments and suggestions of the anonymous reviewers.

Availability of data and material

The SHIBR data set will be available publicly as open access at the following permanent link upon acceptance: https://ardisdataset.github.io/SHIBR/

Compliance with ethical standards

The SHIBR dataset is a subset comprising random samples from the period 1800-1840, which were selected to abide by the GDPR (the European General Data Protection Regulation) law and thus are made open access.

Conflict of interest: Johan Hall is an employee at the company *Arkiv Digital AD AB* (Sweden), the provider of this dataset. Agrin Hilmkil is an employee at the company Peltarion AB (Sweden). The rest of the authors declare that they have no conflict of interest.

References

[1] H. Balk and A. Conteh, IMPACT: centre of competence in text digitisation. In: Proceedings of the 2011 Workshop on Historical Document Imaging and Processing (pp. 155-160), 2011.

[2] H. Balk, Poor access to digitised historical texts: the solutions of the IMPACT project. In: Proceedings of The Third Workshop on Analytics for Noisy Unstructured Text Data (pp. 1-1), 2009.

[3] M. Krystyna and A.H. Qasem, Digitizing the Historical Periodical Collection at the Al-Aqsa Mosque Library in East Jerusalem. In: Proceedings IFLA World Library and Information Congress, Milan, Italy, August 24, 2009.

[4] Z. Zakariah, N. Janom, N.H. Arshad, S.S. Salleh, and S.R.S Aris, Crowdsourcing: The Trend of Prior Studies. In: Proceedings of the 2014 4th International Conference on Artificial Intelligence with Applications in Engineering and Technology (ICAIET'14). IEEE Computer Society, USA, 129–133, 2014. DOI: https://doi.org/10.1109/ICAIET.2014.30.

[5] C. Clausner, J. Hayes, and A. Antonacopoulos, Crowdsourcing Historical Tabular Data: 1961 Census of England and Wales. In: Proceedings of the 5th International Workshop on Historical Document Imaging and Processing (HIP'19). Association for Computing Machinery, New York, NY, USA, 42–47, 2019. DOI: https://doi.org/10.1145/3352631.3352643.

[6] H. Kusetogullari, A. Yavariabdi, A. Cheddad, et al. ARDIS: a Swedish historical handwritten digit dataset. Neural Comput & Applic (2019). DOI: https://doi.org/10.1007/s00521-019-04163-3.

[7] A. Sanchez, P.D. Suarez, C.A.B. Mello, A.L.I. Oliveira and V.M.O. Alves, Text Line Segmentation in Images of Handwritten Historical Documents. In: Proceedings of the 2008 First Workshops on Image Processing Theory, Tools and Applications, Sousse, pp. 1-6, 2008.

[8] K. Zagoris, I. Pratikakis and B. Gatos, Unsupervised Word Spotting in Historical Handwritten Document Images Using Document-Oriented Local Features. IEEE Transactions on Image Processing, vol. 26, no. 8, pp. 4032-4041, Aug. 2017, doi: 10.1109/TIP.2017.2700721.

[9] C. Djeddi, S. Al-Maadeed, A. Gattal, I. Siddiqi, A. Ennaji, HE. Abed, ICFHR2016 competition on multi-script writer demographics classification using "QUWI" database. In: Proceedings of the IEEE international conference on frontiers in handwriting recognition, pp. 602–606, 2016.

[10] S. Ahlawat, A. Choudhary, Hybrid CNN-SVM Classifier for Handwritten Digit Recognition. Procedia Computer Science, Volume 167, pp. 2554-2560, 2020.

[11] R. Alaasam, B. Kurar, M. Kassis and J. El-Sana, Experiment study on utilizing convolutional neural networks to recognize historical Arabic handwritten text. In: Proceedings of the *2017 1st International Workshop on Arabic Script Analysis and Recognition (ASAR)*, Nancy, 2017, pp. 124-128.

[12] F. C. Ribas, L. S. Oliveira, A. S. Britto, R. Sabourin, Handwritten digit segmentation: a comparative study. Int J Doc Anal Recognit, vol. 16, pp. 567–578, 2013.

[13] K. Ntirogiannis, B. Gatos, I. Pratikakis, A combined approach for the binarization of handwritten document images, Pattern Recognition Letters, vol. 35, pp. 3-15, 2014.

[14] D.J. Kennard, A.M. Kent, W.A. Barrett, Linking the past: discovering historical social networks from documents and linking to a genealogical database. In: Proceedings of the 2011 Workshop on Historical Document Imaging and Processing (HIP 2011), New York, USA, 2011, pp. 43–50.

[15] D.W. Embley, S. Machado, T. Packer, J. Park, A. Zitzelberger, S.W. Liddle, N. Tate, D.W. Lonsdale, Enabling search for facts and implied facts in historical documents. In: Proceedings 2011 Workshop on Historical Document Imaging and Processing (HIP 2011), New York, USA, 2011, pp. 59–66.

[16] S. Athenikos, WikiPhiloSofia and PanAnthropon: extraction and visualization of facts, relations, and networks for a digital humanities knowledge portal. In: Proceedings of the 20th ACM Conference Hypertext and Hypermedia (Hypertext 2009), Torino, Italy, 2009.

[17] The Washington Database, Retrieved on 2020-06-20, from: http://www.fki.inf.unibe.ch/databases/iam-historical-document-database/washington-database.

[18] G. Washington, George Washington Papers, Series 2, Letterbooks 1754 to 1799: Letterbook 1- Dec. 25, 1755. [Manuscript/Mixed Material] Retrieved from the Library of Congress, https://www.loc.gov/item/mgw2.001/

[19] R. Sarkar, N. Das, S. Basu, et al. CMATERdb1: a database of unconstrained handwritten Bangla and Bangla–English mixed script document image. IJDAR, vol. 15, pp. 71–83, 2012.

[20] Handwritten Keyword Spotting Competition (H-KWS /ICFHR 2016), Retrieved on 2020-06-20, from: https://www.prhlt.upv.es/contests/icfhr2016-kws/data.html

[21] ICFHR2016 Competitions, Retrieved on 2020-06-05, from: http://www.nlpr.ia.ac.cn/icfhr2016/competitions.htm

[22] The IAM Handwriting Database, Retrieved on 2020-06-20, from: http://www.iam.unibe.ch/fki/databases/iam-handwriting-database

[23] U. Marti and H. Bunke, The IAM-database: An English Sentence Database for Off-line Handwriting Recognition. Int. Journal on Document Analysis and Recognition, vol. 5, pp. 39 - 46, 2002.

[24] M. Kassis, VML-HD: The Historical Arabic Documents Dataset for Recognition Systems (VML-HD). 1, ID: VML-HD1, URL: http://tc11.cvc.uab.es/datasets/VML-HD_1.

[25] W. Pantke, M. Dennhardt, D. Fecker, V. Märgner and T. Fingscheidt, An Historical Handwritten Arabic Dataset for Segmentation-Free Word Spotting - HADARA80P. In: Proceedings of the 14th International Conference on Frontiers in Handwriting Recognition, Heraklion, 2014, pp. 15-20, doi: 10.1109/ICFHR.2014.11.

[26] B. Kiessling, D.S. Ben Ezra, M.T. Miller, BADAM, A public dataset for baseline detection in Arabic-script manuscripts. In Proceedings of the 5th International Workshop on Historical Document Imaging and Processing (HIP'19), ACM, 13-18. DOI: https://doi.org/10.1145/3352631.3352648.

[27] The ESPOSALLES Database, Retrieved on 2020-06-20, from: http://dag.cvc.uab.es/the-esposalles-database/.

[28] V. Romero, A. Fornés, N. Serrano, J.A. Sánchez, A.H. Toselli, V. Frinken, E. Vidal, J. Lladós. The ESPOSALLES Database: An Ancient Marriage License Corpus for Off-line Handwriting Recognition. Pattern Recognition, vol. 46, pp. 1658–1669, 2013.

[29] The IFN/ENIT-database, Retrieved on 2020-06-20, from: http://www.ifnenit.com/download.htm.

[30] R. Hussain, A. Raza, I. Siddiqi, et al. A comprehensive survey of handwritten document benchmarks: structure, usage and evaluation". J Image Video Proc., 46, 2015.

[31] T. Rath, R. Manmatha, Features for word spotting in historical manuscripts. In: Proceedings of the 7th International Conference on Document Analysis and Recognition (ICDAR), pp. 218–222 (2003).

[32] T. Mondal, N. Ragot, J.-Y. Ramel, U. Pal, Performance evaluation of DTW and its variants for word spotting in degraded documents. In: Proceedings of the 13th International Conference on Document Analysis and Recognition (ICDAR), 2015, pp. 1141–1145

[33] A. Bhardwaj, S. Setlur, and V. Govindaraju, Keyword spotting techniques for Sanskrit documents. In: Lecture Notes in Artificial Intelligence 5402, G. Huet, A. Kulkarni, and P. Scharf, Eds. Berlin, Germany: Springer-Verlag, 2009, pp. 403–416

[34] E. Ataer, P. Duygulu, Retrieval of ottoman documents. In: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval, pp. 155–162 (2006).

[35] I. Rabaev, I. Dinstein, J. El-Sana, K. Kedem, Segmentation-Free Keyword Retrieval in Historical Document Images. In: A. Campilho, M. Kamel (eds) Image Analysis and Recognition. ICIAR 2014. Lecture Notes in Computer Science, vol 8814, 2014, Springer.

[36] Y. Leydier, F. Lebourgeois and H. Emptoz, Text search for medieval manuscript images. Pattern Recognition, vol. 40, pp. 3552-3567, 2007.

[37] V. Mane and L. Ragha, Handwritten character recognition using elastic matching and PCA. In: Proceedings of the Int. Conf. Adv. Comput., Commun. Control, pp. 410–415, 2009.

[38] Y. Lu and C. L. Tan, Word Searching in Document Images using Word Portion Matching. In: Proceedings of the International Workshop on Document Analysis Systems (DAS 2002), Springer-Verlag Berlin Heidelberg, LNCS 2423, pp. 319–328, 2002.

[39] A. Fischer, A. Keller, V. Frinken, H. Bunke, (2010). HMM-based word spotting in handwritten documents using subword models. In: Proceedings of the 20th International conference on pattern recognition (ICPR), IEEE, pp. 3416–3419, 2010.

[40] A. L. Bianne-Bernard, F. Menasri, R. H. Mohamad, C. Mokbel, C. Kermorvant, L. Likforman-Sulem, Dynamic and contextual information in HMM modeling for handwritten word recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 33 (10) (2011) 2066–2080.

[41] A. Ahmad, C. Viard-Gaudin, M. Khalid, Lexicon-based word recognition using support vector machine and hidden Markov model. In: Proceedings of the 10th International Conference on Document Analysis and Recognition (ICDAR), pp. 161–165, 2009.

[42] S. Espana-Boquera, M. Castro-Bleda, J. Gorbe-Moya, F. Zamora-Martinez, Improving offline handwritten text recognition with hybrid HMM/ANN models. IEEE Transactions on Pattern Analysis and Machine Intelligence 33 (4) (2011) 1130 767–779.

[43] A. C. Rouhou, Y. K. Kanoun, Hybrid HMM/DNN System for Arabic Handwriting Keyword Spotting. In: Proceedings of the 16th International Conference on Image Analysis and Recognition, Springer, pp 216-227, Canada, August 27-29, 2019. DOI: 10.1007/978-3-030-27202-9_19.

[44] M. W. Sagheer, N. Nobile, C. L. He and C. Y. Suen, A Novel Handwritten Urdu Word Spotting Based on Connected Components Analysis. In: Proceedings of the 20th International Conference on Pattern Recognition, Istanbul, pp. 2013-2016, 2010. Doi: 10.1109/ICPR.2010.496.

[45] J. Almazán, A. Gordo, A. Fornés, E. Valveny, Handwritten Word Spotting with Corrected Attributes. In: Proceedings of the IEEE International Conference on Computer Vision, Sydney, Australia. pp.1017-1024, 2013. DOI: 10.1109/ICCV.2013.130.

[46] V. Frinken, A. Fischer, R. Manmatha, H. Bunke, A novel word spotting method based on recurrent neural networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, no. 2, 211–224, 2012.

[47] P. Krishnan, C.V. Jawahar, HWNet v2: an efficient word image representation for handwritten documents. IJDAR 22, 387–405 (2019).

[48] S. Sudholt and G. A. Fink, PHOCNet: A Deep Convolutional Neural Network for Word Spotting in Handwritten Documents. In: Proceedings of the 15th International Conference on Frontiers in Handwriting Recognition (ICFHR), Shenzhen, pp. 277-282, 2016. DOI: 10.1109/ICFHR.2016.0060.

[49] R. Ahmed, W. G. Al-Khatib, S. Mahmoud., A Survey on handwritten documents word spotting. Int J Multimed Info Retr, vol. 6, pp. 31–47, 2017.

[50] A.A.A. Ali, M. Suresha, Survey on Segmentation and Recognition of Handwritten Arabic Script. SN COMPUT. SCI. 1, 192 (2020).

[51] T. Rath, R. Manmatha, Word spotting for historical documents. IJDAR, 9(2–4), 139–152, 2007.

[52] A. Murugappan, B. Ramachandran, P. Dhavachelvan, A survey of keyword spotting techniques for printed document images. Artif. Intell. Rev., 35 (2) (2011), pp. 119-136

[53] M. Boualam, G. Khaissidi, M. Mrabti, Y. Elfakir, An overview on handwritten documents word spotting. In: Proceedings of the International Conference on Wireless Technologies, Embedded and Intelligent Systems (WITS), 3-4 April 2019.

[54] S. Ren, K. He, R. Girshick and J. Sun, Faster R-CNN: Towards real-time object detection with region proposal networks. In: C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama and R. Garnett (eds.), Advances in Neural Information Processing Systems 28 (Curran Associates, Inc.) pp. 91–99, 2015.

[55] T. Wilkinson, J. Lindström and A. Brun, Neural word search in historical manuscript collections. arXiv preprint arXiv:1812.02771 (2018).

[56] S.B. Hatton, History, Kinship, Identity, and Technology: Toward Answering the Question “What Is (Family) Genealogy?”. Genealogy. 2019; 3(1):2. DOI: 10.3390/genealogy3010002.

[57] K. Abildgren, Mining archival genealogy databases to gain new insights into broader historical issues. Digital Library Perspectives, Vol. 35 No. 3/4, pp. 259-270, 2019. DOI: 10.1108/DLP-07-2019-0025.

[58] Z. Zhu, Content mining and visualization of traditional genealogies of China – Deployed on the genealogy of Wu’s in Gaoqian, Zhejiang. In: Proceedings of the iConference 2020 Sustainable Digital Communities Proceedings. March 23 – 27, 2020, Borås, Sweden.

[59] M. Wojciechowski, M. Zakrzewicz, Dataset Filtering Techniques in Constraint-Based Frequent Pattern Mining. In: Hand D.J., Adams N.M., Bolton R.J. (eds) Pattern Detection and Discovery. Lecture Notes in Computer Science, vol 2447, 2002. Springer, Berlin, Heidelberg.

[60] Statistiska Centralbyrån [National Central Bureau of Statistics]. (1969). Historical Statistics of Sweden: Part 1. Population 1720-1967, Stockholm (2nd edition). Available from: http://share.scb.se/OV9993/Data/Historisk%20statistik/Historisk%20statistik%20f%C3%B6r%20Sverige%201700-1900-tal/Del1-Befolkning-1720-1967.pdf

[61] A.P. Giotis, G. Sfikas, B. Gatos, and C. Nikou, A survey of document image word spotting techniques. Pattern Recognition 68 (2017): 310-332.

[62] A. Cheddad, Towards query by text example for pattern spotting in historical documents. In: Proceedings of the 7th International Conference on Computer Science and Information Technology (CSIT), 13-14 July 2016 Amman, Jordan, pp. 1-6, doi: 10.1109/CSIT.2016.7549479.