Handwriting Text Recognition & Word Spotting Techniques to Build Individual-level Historical Demographic Databases. The Barcelona Case.

Joana Maria Pujadas-Mora*, Alícia Fornés#, Josep Lladós#, Miquel Valls*, Gabriel Brea*

*** Center for Demographic Studies (CED)**

 # Computer Vision Center (CVC

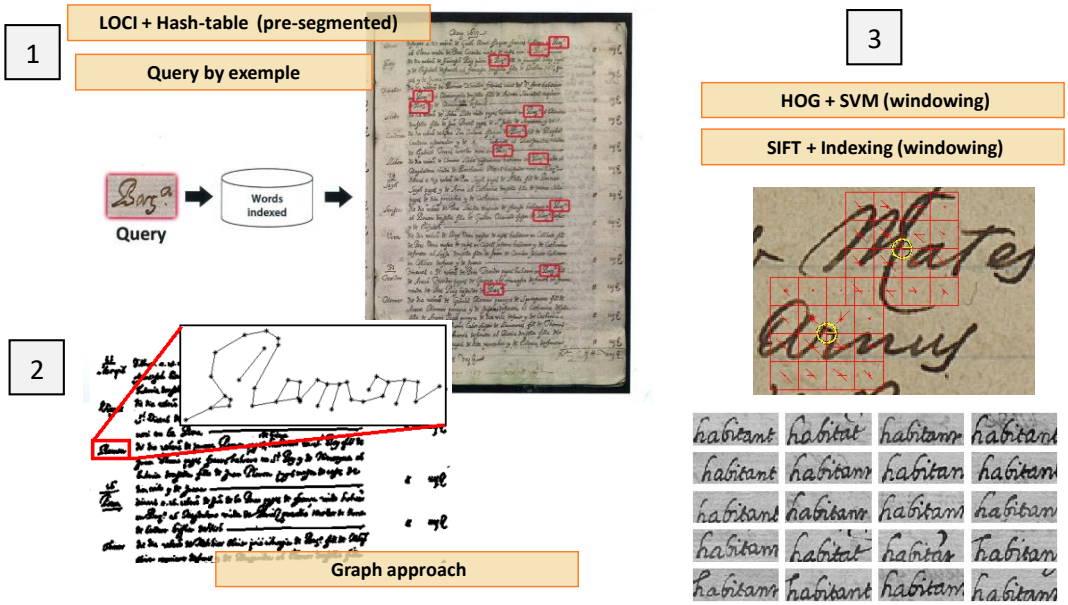
Nowadays, one of the great challenges of Historical Demography is integrating handwriting recognition techniques into data collection of primary sources as a way of being part of the Big Data revolution (Pujadas-Mora et al., 2016). This integration would make possible to reduce the time of data collection and processing large collections of documents and would offer ever-increasing arrays of information. Moreover, this process, also, fits with the gradual introduction of information technology occurring since decades in Humanities, the more recent massive campaigns of digitization of historical sources, which have become customary and the important progress in document image analysis and recognition techniques. In particular, successful adoption of deep learning to handwritten text recognition (HTR) and key word spotting (KWS) has been developed (Toledo et al., 2017). In this times, these techniques are moving towards 'Document Understanding' rather than pure transcription in order to narrow the semantic gap regarding the interpretation of the contents, which is extremely useful to build databases automatically, and more specifically demographic databases. However, it is still a challenge due to the necessity of not only transcribing (recognizing) the original sources, but also to associate the transcribed words to the corresponding semantic classes (corresponding fields of a database).

The aim of the paper is to describe the main document image analysis techniques that have been developed for extracting the information from handwritten demographic sources in order to create the *Barcelona Historical Marriage Database* (BHMD) within the Advanced Grant project 'Five Centuries of Marriages' (IP: Cabré, A.) funded by the European Research Council (http://dag.cvc.uab.es/infoesposalles/) and the *Baix Llobregat Demographic Database* (*BALL*) inside the projects: 'Tools and procedures for the large scale digitization of historical sources of population' (IP: Lladós, J.; Esteve, A.; Pujadas-Mora, J.M.) and 'Networks. Technology and citizen innovation for building historical social networks to understand the demographic past' (IP: Fornés, A.; Pujadas-Mora, J.M.) funded both by RecerCaixa program – Obra Social "la Caixa" (http://dag.cvc.uab.es/xarxes/). The BHMD brings together the marriage licenses recorded at the *Llibres d'Esposalles* covering the Diocese of Barcelona (formed by 250 parishes in 1900) from 1451 to 1905, accounting for more than 600,000 marriages (Pujadas-Mora et al., 2018; Brea-Martínez & Pujadas-Mora, 2018). The BALL database is an ongoing database containing individual census data from the Catalan county of *Baix Llobregat* (Barcelona, Spain), up to now, of 9 different municipalities for the nineteenth and twentieth centuries gathering more than 220,000 individual observations (Pujadas-Mora, et al., 2019).

The specific applied techniques of document analysis in those projects are the Key Word Spotting and the Handwritten Text Recognition. Key Word Spotting is used to locate a particular queried word in document/collection, without the explicit transcription of all the contents of the document (Rusiñol et al. 2011). This technique turns out to be more suitable when a document does not have a clear internal structure or when the handwriting style is new to the system. In this way, it can be used to search and index the contents of the documents. In our

projects, two approaches of word spotting have been employed. A first one is based on a structural and learning-free method (Riba et al., 2017), suitable for searching information in collections where there is no training data at all. This means that it is not necessary to label data previously to train the system and it can be applied to any kind of documentation with any language. In this case, the query is a word image, so the system retrieves those documents containing word images with a similar shape appearance. This approach is known as namely query-by-example word spotting. There are several approaches in this family (Lladós et al; 2012), and some of them are shown in Figure 1. The first one is based on hash-tables that index parts of words (subwords or characters). The second one is based on graphs, which consists on representing the handwritten strokes in a word using a graph, and then applying graph matching techniques to find similar graphs in the documents. The third technique uses gradients (HOG or SIFT descriptors) to characterize the words.

Figure 1: Approaches in query-by-example word spotting apply in 5CofM, TOOLS and NETWORKS projects.



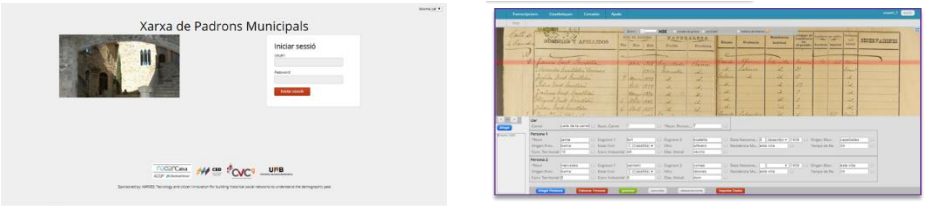
Source: Authors' own elaboration

The second word spotting approach corresponds to a statistical and learning-based method (Almazán et al., 2014), appropriate for searching words in documents where training data is available. This technique is called query-by-string word spotting. In this case, the system learns the correspondence between the letters/words and its corresponding shape appearance (as a word image). This method shows better results because the system can learn the appearance of the text because every writer has their own handwriting style, which evolves with the time. Moreover, this method can perform searches using query by example and query by string. However, the data needs to be labelled to train the system. Integrating both approaches, query-by-example and query-by-string, we have developed and application called *e-crowds* for a mobile platform for browsing and searching into the marriage licenses books. The first approach consists on clicking on a word on a given image, that is query-by-examples and the second one on typing a word in a textual browser, which is query by string (Riba et al., 2014).

To train the system we have followed a strategy based on including the human in the loop. In this way, our research started with the development of a *bimodal crowdsourcing* platform (Fornés et al., 2011). This application integrates two point of views: the semantic information

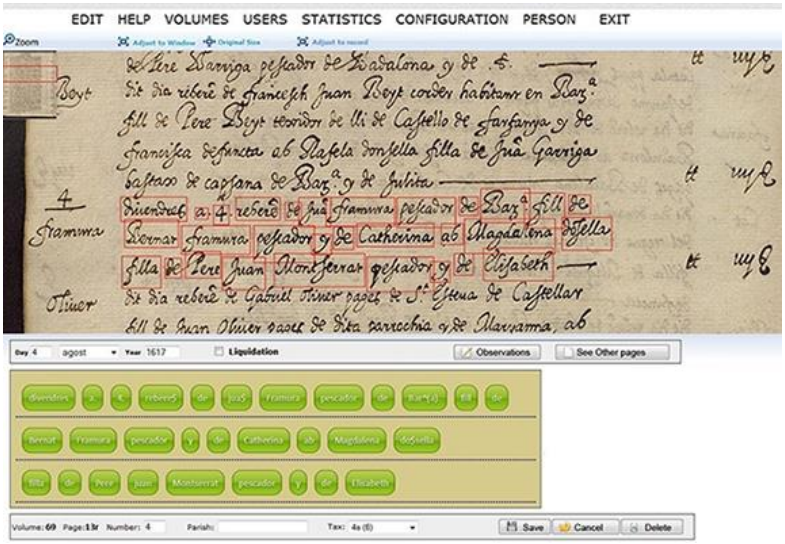
and the ground-truthing for document analysis. The semantic information of the sources is transcribed manually into forms using a friendly environment (see Figure 2). In this way, the graphical user interface of our web-based platform offers a data entry tool which integrates both the original source and the data form in the same view. The ground truth data serves to train or validate the document analysis algorithms required to transcribe (semi) automatically the manuscript demographic sources (Thorvaldsen et al., 2015). This training consists in bounding boxes of words or lines, and annotated data related to their properties which is known as labeling data (see Figure 3).

Figure 2. Crowdsourcing platform for TOOLS and NETWORKS projects:



Source: Authors' own elaboration. http://dagapp.cvc.uab.es/PadronsXarxes/

Figure 3. 'Ground truth' module for the 5CofM project.



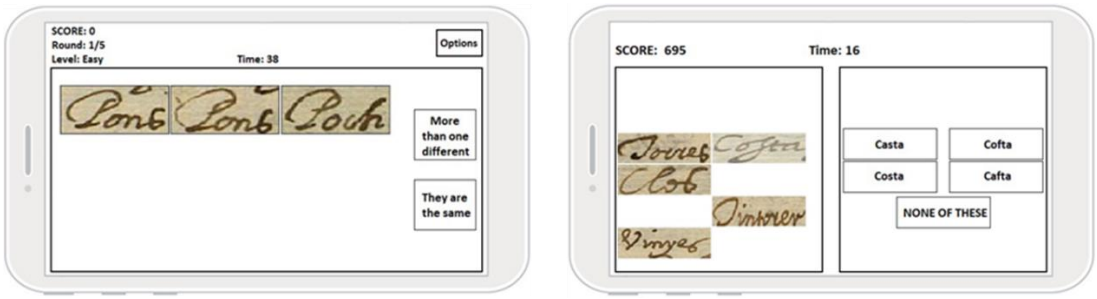
Source: Thorvaldsen, G.; Pujadas-Mora, J.M.; Andersen, T.; Eikvil, L.; Lladós, J.; Fornés, A.; Cabré, A (2015)

When the document is legible, not degraded and there is enough training data of a particular handwriting style, then a handwriting recognition system could be properly trained (Toledo, 2017). In this case, the text image is processed from left to right, and the system outputs a sequence of characters. In this way, once the words are recognized using either handwriting recognition or word spotting techniques, the next step consists in assigning a semantic category, otherwise, the database can not be created. The recognition of semantic categories (e.g. names, places, occupations) is called Name Entity Recognition (NER). Specifically, we propose to recognize the named entities directly over the word image (Toledo et al., 2019; Fornés et al., 2019), which take advantage of the structure of the original sources. This can be seen how it

works in: https://goo.gl/FMTgFy. However not all the demographic documents were recorded in forms as the marriage licenses, but they follow a regular formulation (although some variations may appear). The main idea consists in learning the most common sequences of semantic categories using a Deep neural network (e.g. a name precedes a surname…). Then, the system can benefit from this contextual information when recognizing the semantic labels of a sequence words in a record. Besides, the application of document image analysis techniques needs to overcome the fact that the historical documents suffer of physical degradation, ancient vocabularies, multi-writer styles, variability of the language due to the lack of a standardized version, etc, which reinforce the inclusion of the human.

More recently, gamesourcing experiences (Chen et al., 2018) has been implemented for data annotation and validation of automatic transcription in NETWORKS project (see Figure 4). With this aim, we have developed 2 videogames for mobile devices called word-hunter. They combine automatic transcription and manual validation through gamesourcing, understood as crowdsourcing via gamification. It should be noted that the time spent by the user to validate and correct errors from an automatic transcription is lower than manual transcription from scratch. The first game is a difference game for validation of clusters of words (snippets of documents) proposed by the system. The second game is a match game for validation of automatic transcription. The performance of both games can be watched in: https://goo.gl/4PVupn. It is worth noting that the user engagement in those video games is higher than when using the abovementioned crowdsourcing web-based application and the effort for transcription, annotation or validation is less time consuming.

Figure 4. Android games developed for the NETWORKS project



Source: Chen, J.; Riba, P.; Fornés, A.; Mas, J.; Lladós, J.; Pujadas-Mora, J.M. (2018)

In order to evaluate the performance of these technological developments, quantitative analysis have been performed to obtain their accuracy and detection rates, which will be incorporated in the final version of this paper. These measurements serve as a proof of concept to evaluate if these developments can be applied to transcribe massively.

Quoted references

Almazán, J.; Gordo, A.; Fornés, A.; Valveny, E. (2014) "Word Spotting and Recognition with Embedded Attributes". *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 36 (12): 2552-2566.

Brea-Martínez, G.; Pujadas-Mora, J.M. (2018). “Estimating Long-Term Socioeconomic Inequality in Southern Europe: The Barcelona Area, 1481-1880”. *European Review of Economic History*. DOI: 10.1093/ereh/hey017.

Chen, J.; Riba, P.; Fornés, A.; Mas, J.; Lladós, J.; Pujadas-Mora, J.M. (2018). “Word-Hunter: A Gamesourcing Experience to Validate the Transcription of Historical Manuscripts”. In *16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*:528-533.

Fornés, A.; Lladós, J.; Mas, J.; Pujadas-Mora, J.M.; Cabré, A. (2014). “A bimodal crowdsourcing platform for demographic historical manuscripts”. *DATeCH '14 Proceedings of the first international conference on digital access to textual cultural heritage*. 103-108.

Fornés, A.; Lladós, J.; Pujadas-Mora, J.M. (2019,. forthcoming). “Browsing of the Social Network of the Past: Information Extraction from Population Manuscript Images”. Handwritten Historical Document Analysis, Recognition, and Retrieval - State of the Art and Future Trends. World Scientific, Series in Machine Perception and Artificial Intelligence.

Lladós, J.; Rusiñol, M.; Fornés, A.; Fernández, D.; Dutta, A. (2012). “On the influence of word representations for handwritten word spotting in historical documents”. *International Journal of Pattern Recognition and Artificial Intelligence*, 26:5.

Pujadas-Mora, J.M.; Brea-Martínez, G; Jordà, J.P.; Cabré, A. (2018). “The apple never falls far from the tree: Siblings and intergenerational transmission among farmers and artisans in the Barcelona Area in the 16th and 17th centuries”. *The History of the Family*. 23(4): 533 - 567.

Pujadas-Mora, J.M.; Fornés, A.; Lladós, J.; Cabré, A. (2016). “Bridging the gap between Historical Demography and Computing: Tools for computer-assisted transcription and analysis of demographic sources”. In *The future of historical demography. Upside down and inside out*. Acco, 222-226.

Pujadas-Mora, J.M.; Fornés, A.; Lladós, J.; Brea-Martínez, G; Valls-Fígols, M. (2019) (Forthcoming). “The Baix Llobregat (BALL) Demographic Database, between Historical Demography and Computer Vision (nineteenth – twentieth centuries)”. *Nominative Data for Demographic Research in the East and the West*. Ural Federal University press.

Riba, P.; Lladós, J.; Fornés, A.; Dutta, A. (2017). “Large-scale Graph Indexing using Binary Embeddings of Node Contexts for Information Spotting in Document Image Databases”. *Pattern Recognition Letters*, 87:203-211. (doi:10.1016/j.patrec.2016.06.015).

Riba, P.; Almazán, J.; Fornés, A.; Fernández, D.; Valveny, E.; Lladós, J. (2014). e-Crowds: a mobile platform for browsing and searching in historical demography-related manuscripts. In International Conference on Frontiers in Handwriting Recognition (ICFHR), pp. 228-233)

Rusiñol, M.; Aldavert, D.; Toledo, R.; Llados, J. (2011). “Browsing Heterogeneous Document Collections by a Segmentation-free Word Spotting Method”. *International Conference on Document Analysis and Recognition*. Being, China.

Thorvaldsen, G.; Pujadas-Mora, J.M.; Andersen, T.; Eikvil, L.; Lladós, J.; Fornés, A.; Cabré, A (2015). ‘A tale of two transcriptions. Machine-assisted transcriptions of historical sources’. *Historical Life Course Studies*. 2: 1-19.

Toledo, J.I.; Carbonell, M.; Fornés, A.; Lladós, J. (2019). “Information Extraction from Historical Handwritten Document Images with a Context-aware Neural Model” *Pattern Recognition*, 86:27-36 (doi.org/10.1016/j.patcog.2018.08.020).

Toledo, J.I.; Dey, S.; Fornés, A.; Lladós, J. (2017). “Handwriting Recognition by Attribute embedding and Recurrent Neural Networks”. In *14th International Conference on Document Analysis and Recognition (ICDAR)*:1038-1043.