Context, multimodality, and user collaboration in handwritten text processing: the CoMUN-HaT project

Carlos-D. Martínez-Hinarejos[1], Josep Lladós[2], Alicia Fornés[2], Francisco Casacuberta[1], Lluis de las Heras[2], Joan Mas[2], Moisés Pastor[1], Oriol Ramos[2], Joan-Andreu Sánchez[1], Enrique Vidal[1], and Fernando Vilariño[2]

[1] Pattern Recognition and Human Language Technology Research Center,
Universitat Politècnica de València, Camino Vera s/n, 46022, Valencia, Spain
[2] Centre de Visió per Computador, Dept. Computer Science, Universitat Autònoma de Barcelona,
Edificio O Campus UAB, 08193 Bellaterra (Cerdanyola), Barcelona, Spain

Abstract. Processing of handwritten documents is a task that is of wide interest for many purposes, such as those related to preserve cultural heritage. Handwritten text recognition techniques have been successfully applied during the last decade to obtain transcriptions of handwritten documents, and keyword spotting techniques have been applied for searching specific terms in image collections of handwritten documents. However, results on transcription and indexing are far from perfect. In this framework, the use of new data sources arises as a new paradigm that will allow for a better transcription and indexing of handwritten documents. Three main different data sources could be considered: context of the document (style, writer, historical time, topics,...), multimodal data (representations of the document in a different modality, such as the speech signal of the dictation of the text), and user feedback (corrections, amendments,...). The CoMUN-HaT project aims at the integration of these different data sources into the transcription and indexing task for handwritten documents: the use of context derived from the analysis of the documents, how multimodality can aid the recognition process to obtain more accurate transcriptions (including transcription in a modern version of the language), and integration into a user-in-the-loop assisted text transcription framework. This will be reflected in the construction of a transcription and indexing platform that can be used by both professional and non-professional users, contributing to crowd-sourcing activities to preserve cultural heritage and to obtain an accessible version of the involved corpus.

Keywords: Document processing, context integration, multimodality, user interaction

1 Introduction

In the last decade the interest in automatic processing of historical documents has grown strongly, due to the worldwide mass digitization campaigns for preserving cultural heritage. Documents residing in archives and libraries reflect the identity of the past, so unlocking their contents allows citizens to know the collective and evolving memory of their society. Besides old printed documents, historical documents range from graphical contents like maps or even symbols in stone carvings, to manuscripts written with ink on parchment or paper. In digital libraries many historical manuscripts remain unexploited due to the lack of proper browsing and indexing tools. Scanned document images to be useful, need to be transcribed or indexed. New services emerge with the exploitation of the document contents.

In this context, the emerging field of digital humanities combines the expertise of computer scientists and researchers in humanities and social sciences to provide technological tools that mimic the experiences and procedures of humanities research in the interpretation of historical documents. Scholars in humanities use more and more software that assists in the interpretation of historical manuscripts.

The interpretation of manuscripts is mostly based on Handwritten Text Recognition (HTR). HTR is based on the use of models similar to those employed in automatic speech recognition: optical models that are related to the image of the handwritten text, and language models that are related to the word sequences. HTR has been a research field that has experienced a notable progress in the last decade. Classical techniques are based on Hidden Markov Models (HMM) with n-grams language models [1,2], because they are able to avoid the segmentation of text into characters and words. HMMs are used to estimate the probability for a sequence of feature vectors representing the text images, whereas the concatenation of words into text lines is typically modeled by an n-gram language model. Lately, techniques based on Bidirectional Long Short-Term Memory (BLSTM) Neural Networks (NN) [3] have shown to improve the recognition performance, especially for noisy data. In summary, HMM and NN have defined a solid methodological basis in HTR.

Nevertheless, the complete transcription of manuscripts is not always necessary, since only some parts related to a specific topic are needed. In this sense, the community has made progress on holistic techniques for indexation and information spotting. Word Spotting [4] consists in retrieving all instances of a given word, offering a viable option for searching and browsing information in documents. Word spotting is usually divided into query-by-example (QBE) and query-by-string (QBS) approaches. QBS [5] describes the setup in which the user types the character string to be searched. Although it allows maximum flexibility in terms of vocabulary and handwriting style, this approach requires labeled data to train the system. If no such data is available, QBE can be applied. QBE [6] consists of selecting one or a few words from the documents and the system retrieves all words with a similar shape. Thus, words are treated as visual objects in images, so recognition follows a computer vision paradigm of object detection approach.

However, when applied to historical manuscripts, there are still many challenges to solve in both HTR and word spotting approaches. First, the natural physical degradation of documents due to environmental preservation conditions, chemical effects of the materials (ink, paper) results in very noisy images. Second, HTR methods usually require a large amount of annotated data for training the algorithms. In case of historical documents, these methods require specifically-adapted language models [7] to the specific vocabulary, language, historical time period, etc. For these reasons, annotated data has to be manually created by experts, which turns to be very costly in terms of human effort. Third, writer styles vary a lot not only from one writer to the other, but among time periods or schools. Finally, not only the "shape" of the characters is relevant for the recognition, but the expert knowledge is highly important too. A paleographer "decodes" an old text using his/her expert knowledge on the time period, the domain or some historical facts that can help to understand the contents. In other cases, the analogy with other pages of the source, or with other words in the page help the scholar in the recognition. When the problem scales up, and HTR is applied to large scale volume of data, for example for contents extraction from digital archives and libraries, the error rate in HTR risks of being unacceptable. In any case, the results that provide the state of the art HTR models are still far from perfect and new paradigms, independently of the nature of the optical and language models, have to be added to achieve performance improvements qualitatively relevant.

Due to the difficulties of HTR algorithms in large scale historical databases, an emerging trend for the transformation of raw images into interpretable material is the use of crowdsourcing platforms. It allows speeding up the process, splitting tedious tasks in small and collaborative ones. In addition, and depending on the source documents, it is a form of social innovation, engaging the final consumers of the historical assets, and moving their role from a consumer to an active producer one in the generation of new formats of knowledge of the past. Moreover, scholars in humanities can participate in the validation of historical documents, providing the expert knowledge for the particular source being transcribed. In this sense, the use of multimodal interactive user-interfaces has demonstrated to ease the transcription and validation of the documents [8,9,10]. Indeed, the use of these interfaces is not only limited to document transcription, but even for more advanced processing, such as document translation [11]. Moreover, translation can be considered in a broader sense as a possibility of giving the document semantic added value, or, in case of historical texts, provide the contents of the document in a modern version of the language.

In this framework, the CoMUN-HaT (**Co**ntext, **M**ultimodality and **U**ser collaboratio**N** in **Ha**ndwritten **T**ext processing) was developed by the PRHLT [3] and CVC [4] groups as a follow-up of previous projects that have been developed during the last 10 years (iDoc, MITTRAL, and SearchInDocs). CoMUN-HaT is an ambitious step forward regarding these past projects introducing new paradigms in the recognition process (multimodality and context spaces), added value to historical document transcriptions (text modernisation), and new user experiences at two levels: expert and consumer. The use of multimodality and contextual information in transcription and retrieval, integrated into real scenarios involving communities of users, will end up with a proof of concept of new ways of accessing to digital archives.

2 Project description

The project uses as a starting point the hypothesis that by using contextual information, multimodality, and novel human-computer interaction paradigms, the global transcription and search tasks for handwritten documents would become more accurate and easier. As a final result, a proof of concept tool that can be used by both experts and general users would be developed. In particular, it is planned to use documents containing information of people, events and time. With the recognition of this information, experts can reconstruct socio-economic episodes of the past.

With this starting hypothesis, the objectives of the project are the following:

- To make progress in the state of the art of HTR and information spotting methods applied to mass processing of historical documents.
- To include multimodal features in the handwritten text recognition process.
- To allow the use of modern language variants in transcription and indexing of historical documents.
- To define a context model for historical document recognition at different levels, and to validate its efficiency in the improvement of document transcription and indexation.
- To design a virtual desktop as a metaphor for transcription and retrieval support, where the user is included in the process in a natural way with advanced interfaces (sketching, tangible, multimodal).
- To implement a proof of concept in a real use case scenario in the field of digital humanities, that integrates the generated technologies and validates it with communities of users (at expert and general consumer level).

In order to achieve these objectives, different work packages (up to 10, including management and dissemination) and activities form part of the project. Figure 1 shows a basic scheme on how

[3] `http://www.prhlt.upv.es`

[4] `http://www.cvc.uab.es`

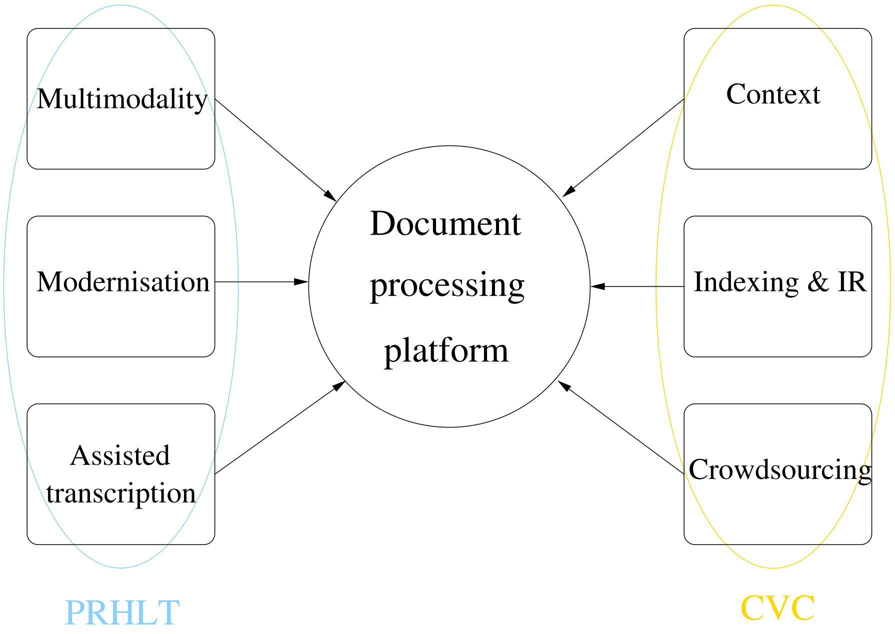

Fig. 1. Basic structure of the activities in the CoMUN-HaT project.

the activities are related and how they converge into the development of a final document processing platform. Project activities started in June 2016 and are expected to finish by December 2018. The following subsections describe more in depth the planned activities.

2.1 Multimodality

This task is related basically to the inclusion of speech as alternative modality for the task of document transcription. The multimodal paradigm aims at combining the different source representations of a given object to obtain its final recognition. In this case, a final sequence of words is represented by a handwritten text image and by a speech signal. The combination of these two signals in order to obtain a better final hypothesis is the final objective of this activity. The obtention of speech signals that contain the dictation of handwritten text is affordable by using a crowdsourcing framework where speech can be acquired from mobile devices (see Subsection 2.6).

In this term, combination can be done at the input level, the search (decoding) process, and at the output level. The asynchronous nature of the two signals (the sequence of feature vectors for text image and speech dictation do not match in length and contents) makes difficult the two first approximations. However, with a proper output representation, output combination is feasible with simpler techniques.

This is the case of the combination of Confusion Networks (CN). CNs are a representation of the alternative hypotheses that a recognizer provides, giving a compact representation of the different sentences without its segmentation. CNs from different recognition process, possibly on different modalities, can be combined to obtain a new CN with hypotheses different from that of the original outputs, and possibly more accurate [12]. In this activity, different combination processes will be defined and implemented to be used with a multimodal corpus and assess their effectiveness with respect to the unimodal alternative.

Another issue to be studied during this activity is the use of natural input units. Different modalities have different basic units for the recognition process, but they refer to the same final object with a common transcription. The use of a common natural unit for all the involved modalities could improve the recognition results. In this activity, the automatic identification and extraction of natural units (sentences) in the involved modalities will be explored, as well as the differences in multimodal recognition results when employing them with respect to those obtained with non-natural units (i.e., lines that do not form complete sentences).

Finally, the use of multimodality would have an impact on the assisted transcription process. The assisted transcription process employs the output of a recognition process with different alternatives (e.g., in the form of a word-graph) in order to offer the user new hypotheses that will reduce the effort in the whole transcription process. Currently, the assisted transcription paradigm is defined on the output of single recognition systems, but for the multimodal case it must be redefined in order to accept the output of the combination models implemented during the project development [13]. Thus, the framework of the assisted transcription process will be redefined to make it accept the output of the multimodal models.

2.2 Context

This task is related to the detection of the context of the document at several levels: historical time, topic, writer, etc. This information is valuable to obtain more specialised models and, consequently, improve the general performance of the transcription task.

The detection of the context would be done at several levels:

1. Logical Layout Analysis: document description from low level entities (text line, graphic entities, images,...) to higher levels of document entities (document title, headers, footnotes, tables,...); probabilistic graphical models would be used in this task o describe models of document according to a layout.
2. Linguistic context models: since n-grams and grammars have shown to be insufficient for solving many ambiguities in the shape of the handwritten strokes in historical and degraded documents, the incorporation of more complex linguistic models will be investigated (e.g., stochastic graph-grammars as bi-dimensional language models), including the extraction of the relevant information in the documents [14] in order to make easier their transcription, annotation and interpretation.
3. Structural context models: a hierarchical attributed graph or hyper-graph is a suitable structure to represent the contextual information, which makes structural and graph theory methods suitable to solve different problems, as locating an information item giving a context (it would be a graph matching or graph traversal problem); this is an emerging field which combines the representational power of graphs [15] with the use of fast and high performance statistical classifiers.
4. Context based on document categorization: the classification and dating of historical manuscripts is often used for sorting the large amounts of material; clusters of documents belonging to the same class (same writer, same structure, same keywords) can be associated to semantic categories, and recognition from one to the other; hence, writer identification and dating methods would be used in order to classify the different documents and for adapting the recognizers.
5. Context based on user-reading models: information about the context-relevant features is intrinsically embedded in the visual search strategies that the experts perform while looking for

cues on the document; the analysis the eye-movements trough eye-tracking will allow putting into correspondence context-related items and to identify the context-relevant features.

6. Semantic knowledge models: knowledge modeling techniques allow to represent the conceptual schema of a specific historical domain; the definition of the ontology and meta-data that represents the contextual information will be used for priming the recognizer in a top-down approach (e.g., type and time period of document will be used to restrict the language, vocabulary, and grammatical structure of the text).

2.3 Modernisation

Transcription of old manuscripts must be available to both specialist and general public. This requires the obtention of transcriptions that are alternative to the literal one (known as "diplomatic" transcription), in order to low down the complexity of the text in vocabulary, grammar structure, and style. This task is usually known as modernisation.

To achieve this aim, a Statistical Machine Translation (SMT) approach would be taken. From several parallel corpora of diplomatic and modern transcriptions, an initial SMT system would be trained and tested in the environment of an interactive-predictive machine translation system. The next step would be the adaptation of the models for sparse training data, since in most of the actual manuscripts that are to be transcribed have a limited amount of draft-transcribed parts, and even fewer have alternative (diplomatic and modern) transcriptions. As a final step, the translation to other languages that allow to widespread the contents of the documents internationally is possible.

2.4 Indexing and Information Retrieval

Many times, the whole transcription is not need to locate an interesting part for a user. The user can make queries in different forms:

- Query by example: the user takes an image containing a word or sequence and looks for other occurrences of that word or sequence; it is specially useful when correct transcription is doubtful for the user.
- Query by string: the user gives a transcription of the word or sequence, which is located in the text (typical word-spotting approach [4]).
- Query by speech: the user dictates the word or sequence that is looking for.

All the modalities may subsume into a query by string by obtaining a preliminar transcription of the image or speech signal, but all the techniques developed in the other activities would allow for a direct search on the image contents.

2.5 Assisted transcription

Thinking of final users, the framework developed on previous activities must be integrated into a model that includes user interaction and assistance [10]. Thus, the main objective in this activity is to develop an interactive assistance model to allow user feedback that includes contextual information, multimodality, and modernization support during the interaction with the user, in order to take advantage of those information sources to reduce the final effort of the user.

The possible forms of integration are the following:

- Context: restriction of user feedback in the assisted transcription based on context information; the context may provide a more accurate recognition of user interaction, and can be used to give the user clues on which feedback is expected.

- Multimodality: user feedback can be obtained in several modalities (e.g., on-line handwriting, speech, keyboard, . . .), and the decoding results for any modality could be improved by taking the available data for the object to be transcribed (e.g., image or speech of the text to be transcribed).
- Modern language feedback: transcription could be done in modern versions of the language, and users could provide feedback in modern language terms instead of using the diplomatic term that is present in the original document; thus, feedback recognition must integrate knowledge of signal sources, diplomatic transcription, and modern versions with the incorporation of the SMT models for modernisation.

2.6 Crowdsourcing

Large-scale transcription of historical manuscripts requires the inclusion of the user in the process. On the one hand, for generating annotated data to train the recognition algorithms, but on the other hand because fully automatic methods are still a challenge and the processes are focused on assisted transcription. In addition, the inclusion of the user involves a social innovation paradigm, where citizens are empowered through online volunteering tasks. The goal of this task is to make progress of the crowd-sourcing platforms developed in previous projects [16] to facilitate the creation of ground-truth of any historical document, using a web-based application that would incorporate multimodal recognition techniques to assists users.

Apart from that, the use of multimodal input allows to widespread the target audience for crowdsourcing to mobile device users, since speech dictation can be used to provide an improvement in the initial transcription of the corpora [17]. Thus, the integration of the speech sources and the consequent development of speech acquisition platforms for mobile devices is another of the objectives of this activity.

3 Expected results

The different activities related to this project have as final objective the development of a document processing platform. This would be reflected into a virtual desktop for final users. Archives and libraries of the future will incorporate novel devices, so tangible interfaces will bring together digital and physical worlds. We can imagine a scholar reading a book placed in a multi-touch table with an aerial camera, so the book is scanned, and integrated into the digital table, where other information are searched and showed to the user. The user will be able to make annotations with digital stylus, crop, select information with pen-based gestures, etc. This virtual desktop would be a metaphor of the innovation in the expert's experience. It will use off-the-shelf devices that incorporate tangible interfaces as HP sprout.

Naturally, this virtual desktop would integrate the data, technologies and software tools generated during the project:

- Data sets with their corresponding ground-truth annotation.
- Software for advanced recognition techniques.
- Multimodal combination methods and software, including detection of natural interaction units.
- Context recognition methods and software.
- Text modernisation SMT systems, including adaptation for sparse data.
- A query system based on image, speech, and text.
- Assisted transcription models and software incorporating context, multimodality, and modernisation.

- Crowdsourcing platforms for document transcription, web and mobile-based.

Figure 2 reflects the different milestones and partial objectives during the development of the project.

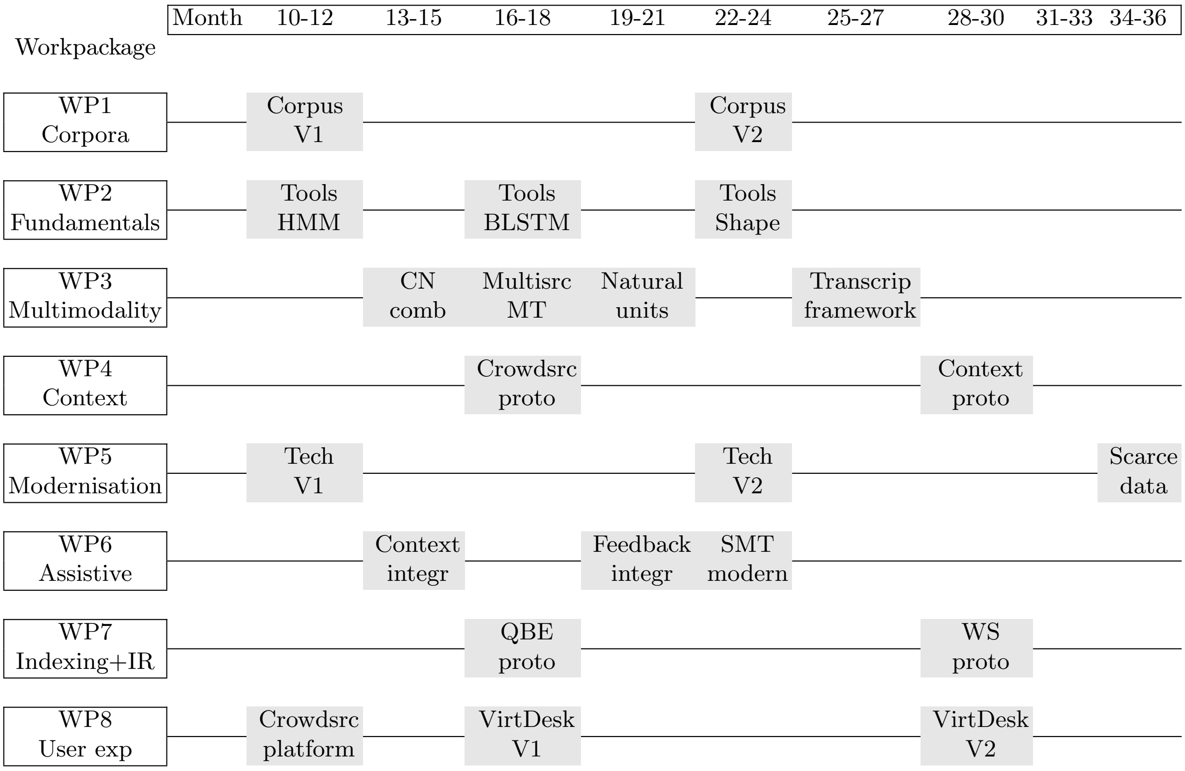

Fig. 2. Milestones during the development of the CoMUN-HaT project.

4 Conclusions

The presented CoMUN-HaT projects shows an ambitious plan to achieve a final platform (virtual desktop) where different interaction modalities (handwritten text, speech, touch, etc.) would cooperate to obtain an enhanced user experience for the task of transcribing and managing handwritten documents. The inclusion of context, modernisation, and multimodality makes this project relevant in its combination of the speech and language technologies with digital image technologies. Thus, the expected results would be of high impact in the future integration of these two areas, that have in common more than it seems at first glance.

Acknowledgments. Work partially supported by project CoMUN-HaT - TIN2015-70924-C2-1-R and TIN2015-70924-C2-2-R (MINECO /FEDER).

References

1. Thomas Plötz and Gernot A. Fink. Markov models for offline handwriting recognition: a survey. *International Journal on Document Analysis and Recognition (IJDAR)*, 12(4):269–298, 2009.
2. Alessandro Vinciarelli, Samy Bengio, and Horst Bunke. Offline recognition of unconstrained handwritten texts using HMMs and statistical language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(6), 0 2004. IDIAP-RR 03-22.
3. Alex Graves, Marcus Liwicki, Santiago Fernández, Roman Bertolami, Horst Bunke, and Jürgen Schmidhuber. A novel connectionist system for unconstrained handwriting recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(5):855–868, May 2009.
4. Volkmar Frinken, Andreas Fischer, R. Manmatha, and Horst Bunke. A novel word spotting method based on recurrent neural networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(2):211–224, February 2012.
5. Jon Almazán, Albert Gordo, Alicia Fornés, and Ernest Valveny. Word spotting and recognition with embedded attributes. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(12):2552–2566, 2014.
6. Marçal Rusiñol, David Aldavert, Ricardo Toledo, and Josep Lladós. Efficient segmentation-free keyword spotting in historical document collections. *Pattern Recogn.*, 48(2):545–555, February 2015.
7. Núria Cirera, Alícia Fornés, Volkmar Frinken, and Josep Lladós. *Hybrid Grammar Language Model for Handwritten Historical Documents Recognition*, pages 117–124. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.
8. Sharon Oviatt, Phil Cohen, Lizhong Wu, John Vergo, Lisbeth Duncan, Bernhard Suhm, Josh Bers, Thomas Holzman, Terry Winograd, James Landay, Jim Larson, and David Ferro. Designing the user interface for multimodal speech and pen-based gesture applications: State-of-the-art systems and future research directions. *Hum.-Comput. Interact.*, 15(4):263–322, December 2000.
9. Vicent Alabau, Carlos-D. Martínez-Hinarejos, Verónica Romero, and Antonio-L. Lagarda. An iterative multimodal framework for the transcription of handwritten historical documents. *Pattern Recognition Letters*, 35:195 – 203, 2014. Frontiers in Handwriting Processing.
10. Alejandro H. Toselli, Verónica Romero, Moisés Pastor, and Enrique Vidal. Multimodal interactive transcription of text images. *Pattern Recogn.*, 43(5):1814–1825, May 2010.
11. Vicent Alabau, Alberto Sanchis, and Francisco Casacuberta. Improving on-line handwritten recognition in interactive machine translation. *Pattern Recogn.*, 47(3):1217–1228, March 2014.
12. Emilio Granell and Carlos D. Martínez-Hinarejos. Combining handwriting and speech recognition for transcribing historical handwritten documents. In *13th International Conference on Document Analysis and Recognition (ICDAR 2015)*, pages 126–130. IEEE Computer Society, 2015.
13. Emilio Granell, Verónica Romero, and Carlos D. Martínez-Hinarejos. An interactive approach with off-line and on-line handwritten text recognition combination for transcribing historical documents. In *12th IAPR International Workshop on Documents Analysis Systems (DAS 2016)*, pages 269–274, 2016.
14. Verónica Romero, Alicia Fornés, Enrique Vidal, and Joan Andreu Sánchez. Using the mggi methodology for category-based language modeling in handwritten marriage licenses books. In *15th International Conference on Frontiers in Handwriting Recognition (ICFHR-2016)*, pages 269–274, 2016.
15. Pau Riba, Josep Lladós, Alicia Fornés, and Anjan Dutta. Large-scale graph indexing using binary embeddings of node contexts for information spotting in document image databases. *Pattern Recognition Letters*, pages –, 2016.
16. Alicia Fornés, Josep Lladós, Joan Mas, Joana Maria Pujades, and Anna Cabré. A bimodal crowdsourcing platform for demographic historical manuscripts. In *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, pages 103–108. ACM, 2014.
17. Emilio Granell and Carlos D. Martínez-Hinarejos. A multimodal crowdsourcing framework for transcribing historical handwritten documents. In *16th ACM Symposium on Document Engineering (DocEng 2016)*, pages 157–163, 2016.