Jornadas de Seguimiento de Proyectos, 2009
Programa Nacional de Tecnologías Informticas

iDoc: Interactive Analysis, Transcription and Translation of Old Text Documents TIN2006-15694-C02

Alfons Juan *
Instituto Tecnológico de Informática
Univ. Politécnica de Valencia

Josep Lladós †
Centro de Visión Por Computador
Univ. Autónoma de Barcelona

Abstract

There are huge historical document collections residing in libraries, museums and archives that are currently being digitised for preservation purposes and to make them available worldwide through large, on-line digital libraries. The main objective, however, is not to simply provide access to raw images of digitised documents, but to annotate them with their real informative content and, in particular, with text transcriptions and, if convenient, text translations too. iDoc aims at developing advanced techniques and interfaces for the analysis, transcription and translation of images of old archive documents, following an interactive-predictive approach. Our hypothesis is that this goal cannot be reliably accomplished by fully automatic techniques; instead, a person-machine collaborative model has to be followed so as to produce accurate document interpretation in a cost-effective way. In order to show our hypothesis, a software tool is being developed and periodically evaluated in terms of usability and profitability. Roughly speaking, our basic goal is to reduce the time needed to manually transcribe/translate a moderately complex text document by at least 25%.

1 Objectives

As indicated above, the main objective of iDoc is to develop advanced techniques and interfaces for the analysis, transcription and translation of images of old archive documents, following an interactive-predictive approach. To achieve this main objective, iDoc has been planned as a coordinated project with two subprojects: iAnaDoc ("Interactive Analysis of Old Archive Documents") and iTransDoc ("Interactive Transcription and Translation of Old Text Documents"). As suggested by their names, iAnaDoc mainly covers the image analysis part, while iTransDoc is primarily devoted to the transcription and translation tasks.

As stated in the proposal, the global objectives of iDoc are:

1. To select and prepare adequate historical document collections
2. To define adequate specifications, benchmarking tasks and evaluation methodologies for the project software
3. To develop appropriate techniques for document image enhancement

*Email: ajuan@dsic.upv.es
†Email: josep@cvc.uab.es

4. To design advanced algorithms for interactive document image recognition
5. To develop novel, interactive machine translation for old-modern language pairs
6. To design advanced, multimodal interfaces to optimise user interaction
7. To develop and assess two software prototypes: a first prototype with basic features and a second prototype with all features included

In order to achieve these objectives, a work plan was designed in which each of them is tackled in a separate work package (WP). Indeed, the work plan consists of 8 WPs, numbered from 0 to 7, and WPs 1 to 7 are in one-to-one correspondence with the above list of global objectives. These WPs and their associated tasks are:

WP	Task	Leader	Oct 06–Sep 07	Oct 07–Sep 08	Oct 08–Sep 09
WP0: Coordination	Coo-org	ITI			
	Coo-net	ITI			
	Coo-mee	ITI			
	Coo-doc	ITI			
	Coo-kno	ITI			
WP1: Corpora preparation and prototype testing	Cor-sel	CVC			
	Cor-dig	CVC			
	Cor-exp	CVC			
	Cor-tra	ITI			
	Cor-man	CVC			
	Cor-tes	ITI			
WP2: Specifications and benchmarking	Spe-fil	CVC			
	Spe-bas	CVC			
	Spe-mt	ITI			
	Spe-adv	ITI			
	Spe-pro	ITI			
WP3: Doc. Image Proc. for Enhacement	Enh-icc	CVC			
	Enh-fie	CVC			
	Enh-gtr	CVC			
WP4: Document Image Recognition	Rec-dla	CVC			
	Rec-p2l	CVC			
	Rec-ord	CVC			
	Rec-str	CVC			
	Rec-grr	CVC			
	Rec-ocr	CVC			
	Rec-htp	ITI			
	Rec-htr	ITI			
	Rec-hts	ITI			
WP5: Interactive Machine Translation	Imt-cor	ITI			
	Imt-mod	ITI			
	Imt-sch	ITI			
	Imt-eva	ITI			
WP6: User Interaction	Usi-spe	ITI			
	Usi-oht	ITI			
	Usi-ske	CVC			
WP7: Software Integration	Sft-net	ITI			
	Sft-1st	CVC			
	Sft-2nd	ITI			

In this report, we only consider the activities carried out before this year. Note that the work planned for this year has been indicated as green squares in task timings.

2 Success level reached

The global objectives are being successfully met, according to the forecasted planning. For each WP (global objective), a Section is included below in which we discuss the results obtained so far, the difficulties found, and the proposals to cope with them.

2.1 Coordination (WP0)

Coo-org, Coo-net, Coo-mee and Coo-doc: Organisation, website, e-mail lists, meetings and documentation. We are using an integrated, web-based project management system, SVN repositories, diverse e-mail lists, etc. However, no strict rules are being applied with respect to document formats, and we simply insist in complying with the acknowledgement normative in the BOE (9-12-2005) on project publications and results. With respect to coordination, a kick-off meeting was held at the beginning of project activities, as planned. Then, regular and extraordinary meetings have been hold at different levels, in accordance with the work plan. It is worth noting here that, on June 2008, we jointly organised an international workshop on project-related topics at Barcelona [6, 11]. Nevertheless, as stated in the proposal, the two groups participating in iDoc have *complementary* expertise in the different scientific areas involved in the project, and thus no special care is needed to coordinate project activities in most tasks. Moreover, a post-doc student from iAnaDoc joined iTransDoc on September 2008, which facilitates integration of iAnaDoc results in the project prototypes.

Coo-kno: Management of knowledge and results. *ITI:* As expected, we have signed an agreement with our Promoter-Observer Entity, BiValDi, for their support in WP1. Moreover, as BiValDi did not have appropriate data for WP5, we tried different options to get such data and, finally, we have recently signed an agreement with the Institut Cambó to make use of its well-known Bernat Metge collection (see WP5). *CVC:* We have also signed a collaboration agreement with ESAGD, which has provide the BORDER-FILES and SANT-ESPERIT datasets. In addition, we have signed a contract with Xerox Research Centre (Grenoble) permitting joint collaborations in task Rec-hts in WP4.

2.2 Corpora preparation and prototype testing (WP1)

Cor-sel: Selection of historical document collections. Two main historical document collections have been selected for their use in the project: GERMANA, from BiValDi, and BORDER-FILES and SANT-ESPERIT, from ESAGD. GERMANA is a 764-page Spanish manuscript entitled *"Noticias y documentos relativos a Doña Germana de Foix, última Reina de Aragón"* and written in 1891 by V. Salvador. BORDER-FILES is composed of 93 linear meters of handwritten and printed documents from 1940 to 1976, about people in the Spanish-French border. There are many other document collections used in different project tasks, such as SANT-ESPERIT: a batch of musical scripts composed of 963 handwritten opus, from which 291 are from anonymous composers.

Cor-dig, Cor-exp, Cor-tra, Cor-man and Cor-tes: Digitisation, expert analysis, manual annotation and prototype testing. *ITI:* GERMANA was carefully scanned by experts from BiValDi at 300dpi in true colours. Following BiValDi recommendations, a palaeography expert was contracted part time on February 2007 so as to work on these tasks. As a result, GERMANA has been analysed and manually annotated in terms of (text) blocks, lines and transcriptions [56]. Also, a first project prototype has been developed following the recommendations made by the expert after intensive testing [71]. *CVC:* BORDER-FILES was already digitised at the beginning of the project. It was partially annotated using an interactive tool in tasks Rec-p2l and Rec-ord of WP4. Furthermore, SANT-ESPERIT was digitised during 2007 and 2008.

2.3 Specifications and benchmarking (WP2)

We decided not to put a lot of effort in specifying file formats and project modules for basic operations, machine translation, advanced interaction, and project prototypes. Regarding project prototypes, we decided not to build them from scratch, but on top of GIMP (GNU Image Manipulation Program), since GIMP gives us for free many desired prototype features. In particular, GIMP is Free Software (GPL), multi-platform, multilingual, and provides: a high-end user interface for image manipulation; a large collection of image conversion drivers and low-level processing routines; an scripting language to automate repetitive tasks; an API for installation of user-defined plug-ins; etc. On the other hand, thanks to experience of our groups on the project topics, we already have many lab prototypes (modules) for specific tasks, both under construction and finished, and thus we decided to work of them within the project framework. An example of this is the interactive machine translation prototype that was developed by our group in the framework of TT2, the European project on machine translation that inspired iDoc.

2.4 Document Image Processing for Enhacement (WP3)

We are mostly using conventional document image processing techniques to facilitate the analysis and recognition tasks in other WPs. However, some novel techniques have been investigated. In particular, the correction of gradient backgrounds, which is especially useful in the case of historical documents with graphical or artistic elements [99, 100].

2.5 Document Image Recognition (WP4)

Rec-dla: Document layout analysis (segmentation) GERMANA is not a particularly difficult task for document layout analysis since most pages only contain nearly calligraphed text written on ruled sheets of well-separated lines. On the contrary, BORDER-FILES is a much more difficult task because of the rich structural diversity of the documents in it. We have been working on original approaches for segmenting the different object categories present in a document: text/graphics separation [122, 119, 120, 118, 116], touching character segmentation, and text segmentation in colour document images.

Rec-p2l and Rec-ord: Physical-to-logical structure mapping, reading order and crosslinking. Our work has focused on three main tasks. First of all, we have developed an interactive tool, based on sketching interfaces, to help in creating and editing the ground-truth regarding the zones and the category of the documents [105]. We have used this tool to partially annotate BORDER-FILES at word-level (see WP6). Secondly, another line of research is trying to combine several of these methods in order to improve their individual performance. To that extent, some alternatives to combine the output of commercial segmentation engines have been proposed and evaluated with promising results [92]. Finally, we have developed a new method for pre-segmentation of document pages that can be used as a pre-processing step in any of the existing methods for layout analysis. This method is currently under evaluation.

Rec-str: Structure classification and matching. We have developed a graph-based representation of the layout that can be serialised into a vector representation. It has been applied to the classification of BORDER-FILES pages. Also, some of our recent work on graph theory [88, 87, 89, 91, 90, 131] allow us to improve the classification and clustering of graph-based representations and could be used in conjunction with this new graph-based representation.

Rec-grr: Graphics recognition. We have intensively worked on this topic and, indeed, we have achieved some international visibility [13, 15]. More specifically, we have been working on three research lines:

1. *Shape description and evaluation.* Shape features play an important role in graphics recognition. We have proposed a framework to review and evaluate the general performance of several shape descriptors [107, 129]. It has been applied to the case of symbol recognition in the third international contest on symbol recognition, which was organised by our group [15, 130]. Also, we have developed descriptors for musical score recognition [85, 86, 93, 94, 96, 95] and we have proposed different ways of combining multiresolution descriptors at different scales, to obtain a single shape representation [17].
2. *Symbol spotting.* We have proposed different methods to retrieve, from a large database of technical line-drawings, zones of interest likely to contain an instance of the graphical symbol queried by the user [123, 126, 125, 128]. We have also proposed a method to automatically locate and recognise graphic symbols in low-resolution images [124, 127]. Another work related to the problem of query by shape is described in [83], where a fast image retrieval system is proposed by using a Run Length Encoding-based image comparison algorithm.
3. *Performance evaluation.* A new approach has been proposed for the generation of synthetic graphics documents containing non-isolated symbols in a real context [82]. This approach enables us to automatically generate large, groundtruthed image database for evaluation purposes. Also, an overview on performance evaluation has been presented in [84], as a result of the work undertaken by a working group on this subject. We are now working on signature methods of regions of interest, and the adaptation of information retrieval evaluation measures to graphics recognition.

Rec-ocr: Optical Character Recognition. On the one hand, as planned, we have studied classifier combination schemes, and a new approach has been proposed that, under certain conditions, gives optimal classifier combination [106]. We are now evaluating standard OCRs to select those with the best individual performance, and then apply our new approach to them. On the other hand, we have developed OCRs specifically adapted to complex layouts [117].

Rec-htp: Handwritten text preprocessing and recognition. As indicated in the proposal, we have studied and improved a number of techniques for handwritten text preprocessing and feature extraction [9]. Most of them have been included in the first project prototype (see WP7). On the other hand, in accordance with the work plan, we have been studying possible improvements to both, HMM and language modelling. Regarding HMM modelling, we have proposed new, Bernoulli mixture-based emission probability functions in which the usual geometric transformations of text can be explicitly modelled [58, 33]. With respect to language modelling, we have studied the combination of n-grams and stochastic incontextual grammars in [75, 60]. Moreover, we are currently developing new approaches for writer identification in SANT-ESPERIT [98] (see WP1).

Rec-hts: Search engines for handwritten text recognition. As planned, we have investigating fast search procedures so as to achieve (nearly) real-time handwritten text decoding. We started from a Viterbi-based search method for text image and transcription alignment [61]. Then, it was replaced by a more efficient method based on word graphs, which works in real-time [62, 76]. Also as indicated in the proposal, we have been working on the estimation of confidence measures. First, our conventional word graph-based method has been integrated in the first project prototype (see WP7). Second, we have explored other possibilities for confidence estimation such as support vector machines and maximum entropy models [29]. Finally, we have carried out some work on *handwritten word spotting,* that is, the task of detecting keywords in handwritten document collections without the explicit use of handwriting recognition techniques. In this task, we have obtained improved results at both, feature and statistical modelling level [102, 101, 109, 108, 114, 115, 113].

2.6 Interactive Machine Translation (WP5)

Imt-cor: Corpus preprocessing. One of the major challenges in this WP is the compilation of a large bilingual parallel corpus having Latin as a source language. As said in Coo-kno (WP0), BiValDi did not have an appropriate corpus for this WP, and thus we tried to find it by other means. We first searched on the Internet, where the best option found was two biblical electronic documents, Novavulgata and bibMaryland, which were preprocessed and aligned at sentence level. However, after a detailed analysis of their characteristics, we realized that they were also not suitable for this WP. We then contacted with the Institut Cambó so as to have access to their large Bernat Metge collection, which includes more than 300 books in Latin together with high-quality translations into Catalan. After signing an agreement with the Institut Cambó, we have recently started to work on some books and results are expected soon. Apart from using the Bernat Metge collection, which surely meet our large data requirements, we have also developed a web crawler to automatically gather bilingual documents from Wikipedia, as well as Latin-Spanish parallel corpora from the Internet [74]. It must be noted that the lack of a large corpus involving Latin has not had negative effects on the other tasks in this WP. They are being carried out normally, using large corpus not involving Latin, but we are almost sure that the results can be extended to Latin in the next months.

Imt-mod: Training and evaluation of translation models The work developed exceeds our initial plan. It can be divided into the following three subtasks:

1. *Phrase-based models.* On the one hand, we have been working on the development and estimation of statistical generative models with promising results [23]. On the other hand, statistical heuristic models were also investigated using the formalism of stochastic inversion transduction grammars (SITG). Both models, as well as many other phrase-based models, do no scale well with large corpora, and hence solutions are needed to cope with the computational bottleneck. We have been working on this issue first, by developing a general framework in which different scaling techniques are combined [53], and second, by using advanced preprocessing techniques of phrase tables [65].
2. *Translation using phrase-based systems.* The usual practice of using log-linear combinations of translation models has been analysed theoretically, and it has been shown that these log-linear combinations can be seen as the application of different loss functions [1]. Also, we have explored how to optimise the parameters of a combination, and a new method based on support vector machines has been proposed which outperforms in some cases state-of-the-art techniques [40]. Another important issue about translation using phrase-based models is the limited matching capability of source-to-source and source-to-target phrases, which we have studied and improved in [73] and [54], respectively. Apart from the above, we have also studied how to include morpho-syntactic information [70], the application of STIG to obtain phrase segments [69, 68], the use of phrase-based models as POS taggers [31], the adaptation of Adaboost to machine translation [45], and the enriching of phrase-based models with trigger models [41].
3. *Translation using other models.* An indirect way of bettering the quality of phrase-based models is to improve word alignments, which we investigated in [27] and [39]. On the other hand, we did some progress in our long-standing research line on stochastic finite-state transducers for machine translation [4, 20, 34]. The latest developments in this line include: the implementation of efficient training and search procedures, the integration of linguistic information, and the incorporation of statistical phrases [36, 37, 35].

Imt-sch and Imt-eva: Search engines and evaluation in interactive translation. Most of the work discussed in the preceding task also applies to this task. On the other hand, as planned, we have studied the use of word-level confidence measures in interactive machine translation [63]. Also, we have extended the use of these measures to stochastic parsing [24], for which we recently proposed optimised techniques [25, 32]. Finally, in [66], more sophisticated human-computer interaction schemes have been proposed to enhance the ergonomic conditions and increase user productivity.

2.7 User Interaction (WP6)

Usi-spe: Speech recognition-based interaction. An improved version of our speech recogniser, iATROS, has been developed and adapted to both speech, and off/on-line handwritten text recognition [50]. In accordance with the proposal, we have considered the use of speech in interactive applications [46, 47], including computer-assisted speech transcription [57]. Also, this work entailed some research on related tasks such as dialog processing of spoken dialogs [7, 51, 52, 72] and language and speaker adaptation [43, 49, 48]. Moreover, in connection with WP5, we have studied new models for speech confidence estimation [30], as well as the more general problem of speech-to-speech translation [19, 21, 55]

Usi-oht: On-line handwritten text recognition. As planned, an on-line handwritten text recognition engine was developed and tested as an additional input modality to improve the interactive performance. The engine itself ranked third in the an international competition on Online Tamil Handwritten Character Recognition Competition [78]. To test it as an additional input modality, we bought a Wacom Cintiq touchscreen and a web-based demonstration prototype was built on top of it for computer-assisted handwritten text transcription [59, 77]. As expected, this modality results in a friendly, highly productive user interface, and thus it will be integrated in the second project prototype. All the work done in this task was carried out in close coordination with that in task Rec-hts (WP4).

Usi-ske: Sketching. Several web and sketch-based interfaces have been developed: to create, edit and link document information [103, 104]; to extract layout information (see WP4); and to retrieval by sketch. Some of them have been applied to the BORDER-FILES dataset as well as in on-line correction systems [109, 111, 112]. Also, similar sketch-based retrieval techniques have been applied to object and image retrieval [80, 81]

2.8 Software Integration (WP7)

As discussed in WP2, we decided not to build project prototypes from scratch, but on top of GIMP (GNU Image Manipulation Program). Although the development of the first prototype was planned from the very beginning of the iDoc, it was not until October 2007 that we started, after hiring three computer scientists. We have just released the first project prototype [56]. It has been implemented as a set of GIMP plug-ins and includes many of the know-how and tools obtained in WP4. Also, it has been successfully used by our palaeography expert to complete the annotation of GERMANA [71]. We are currently working on the second prototype, which is being developed from the first, by integration of multimodal interaction modules developed independently in WP6.

3 Results indicators

3.1 First subproject: iTransDoc (TIN2006-15694-C02-01)

Goal achievement iTransDoc-specific, as stated in the proposal, are being successfully met up to a great extent, say, 70%. Basically, there are only two tasks delayed: corpus preprocessing for interactive machine translation, and the second prototype. On the one hand, as discussed in section 2.6, we have recently started to work on some books from the Bernat Metge collection and results are expected soon. On the other hand, as indicated in section 2.8, we are currently working on the second prototype, which is being developed from the first, by integration of multimodal interaction modules.. All in all, we think that project activities will finish in time and we will completely achieve project objectives.

Scientific production and relevance of results. iTransDoc has produced 1 book, 7 journal articles and 2 PhD thesis during 2007 and 2008 (see References). Note that articles have been published in relevant international journals such as *Machine Learning, IEEE Signal Proc. Magazine, Pattern Recognition Letters and Speech Communication.* On the other hand, iTransDoc scientific production also includes 61 papers in well-known international conferences such as ICDAR, ICIAR, ICASSP, EMNLP, RANLP, EAMT, LREC and IbPRIA (see References).

Utility of results and their role in economic and social development. iTransDoc results are directly applicable to the main research areas of our group and their associated technology transfer projects. Our most important active project of this kind is i3media (2007-2010): a "tractor" technology project within the Spanish *Programa CENIT-Ingenio 2010,* run through a consortium of 12 main enterprises of the media sector, which also involve 19 research groups, including ours. i3media focuses on the creation and automated management of *intelligent audiovisual content,* so as to facilitate both, content personalisation and interaction with users (see `i3media.barcelonamedia.org`). Our participation in i3media is centred on interactive-predictive machine translation, to transfer and adapt our experience on this technology to i3media-specific needs.

Human Resource Development. iTransDoc has grown from 16 members in the original research team, to 31 members (as of the end of 2008). From them, 13 are PhD: 10 PhDs from the original team; 2 students from the original team that have finished their PhD thesis within iTransDoc; and a post-doc student from the CVC group (iAnaDoc) who has joined us on September 2008 under a three-year contract supported by the Spanish *Juan de la Cierva* program. The other 18 members are mainly PhD students supervised by PhD iTransDoc researchers: 10 are fellowships supported by grants from the Spanish government (6), the Valencian Generalitat (2) and our university (2); 2 are university professors; and 6 are research assistants hired for both iTransDoc (3) and other projects (3). It must be noted that iTransDoc has hired 8 research assistants in total: 4 on February 2007 and 4 on October 2007. From the first 4, 3 left contracts early since they became fellowships of the Spanish government (2) and our university (1), under the framework of iTransDoc. The other assistant from the first 4 is a palaeography expert who still works (part-time) for the project. Regarding the second 4, 3 are young computer scientists who have been following the IARFID Master Program while working for iTransDoc, mainly on the development of project prototypes. One of them, however, left his contract on September 2008 after becoming fellowship of the Spanish government, also under the iTransDoc framework. The remaining assistant from the second 4 also left his contract on September 2008, but he still works for iTransDoc, now as professor of our university.

Collaborations with other European or international groups. Our group maintains active collaborations with many national and international research groups, mainly under the framework of integrated actions (with Portugal and Germany), research networks and projects.

Our most important active Spanish project is MIPRCV (2007 - 2012): a research program funded by the Spanish *Programa Consolider-Ingenio 2010,* which, under the coordination of our group, involves more than 80 highly qualified scientists and engineers from seven research groups and ten different public research institutions (see `miprcv.iti.es`). Generally speaking, its main goal is to show how existing Pattern Recognition and Computer Vision technologies can naturally evolve to help developing advanced multi-modal interactive systems. On the other hand, our group belongs to the thematic networks on *Speech Technologies* (see `www.rthabla.es`) and *Audio-Visual Signal Processing in Advanced Multimodal Interfaces* (see `cyberpc.ugr.es/RT`) where, indeed, the first project prototype was publicly presented in its last meeting on December 2008. Regarding international collaborations, our most recent project is TT2 (2002 - 2005): an FP5 European project in which our group participated together with other 6 partners from the European Union and Canada. TT2 influenced very much the way in which we approach research tasks and inspired iDoc as well as MIPRCV. It also inspired other proposals of European projects, including a new proposal for the FP7 ICT Call 4 on Language-based interaction, in which we have great expectations. On the other hand, we have just been accepted as a member of the PASCAL2 network, a well-known Network of Excellence funded by the European Union (see `www.pascal-network.org`).

Project coordination, development and management. Please see section 2.1.

3.2 Second subproject: iAnaDoc (TIN2006-15694-C02-02)

Goal achievement The main goals proposed for iAnaDoc were the investigation and development of approaches for document contents extraction, mainly focusing on non-textual contents, namely structure and shape. Additionally iAnaDoc proposed to study the inclusion of the user in the annotation of documents using sketching interfaces. In general, the objectives in the above lines have been achieved in a 60% (considering the first two years of the project). To complete the objectives, in the third year the results that have been obtained should be published in relevant journals. In addition, 5 thesis will be presented during 2009.

Scientific production and relevance of results. iAnaDoc has produced 5 publications in terms of books, book chapters and journal articles, as well as 2 PhD thesis (see References). Note that this includes two articles into the highly relevant *Int. J. on Document Analysis and Recognition,* and an article in the well-known *IEEE Trans. on PAMI* journal. Also, iAnaDoc has published the results obtained so far in 52 papers of relevant international conferences such as ICPR, ICDAR, DAS, ICFHR and IbPRIA (see References). As described above, currently 7 papers has been submitted to indexed journals (in 1st and 2nd review), and 5 papers are in preparation to be submitted in the near future to relevant journals.

Utility of results and their role in economic and social development. Some of the results have been evaluated and transferred to Promoter-Observer Entities (EPOs). In particular, a prototype for annotation and information extraction has been developed for the BORDER-FILES collection. A second EPO proposed initially was Icar Vision System. Here, the collaboration in the project has been extended to other projects and contracts, not directly in the field of historical document analysis, but with close topics such as handwriting analysis for signature verification and logo recognition in administrative documents.

Human Resource Development. During the first two years, two students belonging to the research team have finished their PhD thesis [12]. Also, 5 students will finish in 2009.

Collaborations with other European or international groups. Collaborations in the topics of iAnaDoc have been carried out with: INRIA Nancy (France), Universite de La Rochelle (France), University of Bern (Switzerland), and Xerox Research Centre (France).

Project coordination, development and management. Please see section 2.1.

References

iTransDoc books, journal articles & PhD thesis

[1] J. Andrés et al. On the use of different loss functions in statistical pattern recognition applied to machine translation. *Patt. Rec. Lett.*, 29, 2008.

[2] J. Andrés et al. Statistical estimation of rational transducers applied to machine translation. *Applied Artificial Intelligence*, 22(1-2):4–22, 2008.

[3] F. Casacuberta, M. Federico, H. Ney, and E. Vidal. Recent efforts in spoken language translation. *IEEE Signal Proc. Mag.*, 25(3), 2008.

[4] F. Casacuberta and E. Vidal. Learning finite-state models for machine translation. *Machine Learning*, 66(1):69–91, 2007.

[5] J. Civera. *Novel statistical approaches to text classification, machine translation and computer-assisted translation*. PhD thesis, Univ. Politécnica de Valencia, June 2008.

[6] A. Juan and G. Sánchez, editors. *Pattern Recognition in Information Systems*. INSTICC PRESS, Barcelona (Spain), June 2008. ISBN 978-989-8111-42-5 (same as [11]).

[7] C.-D. Martinez et al. Statistical framework for a spanish spoken dialogue corpus. *Speech Comm.*, 50:992–1008, 2008.

[8] D. Ortiz et al. The scaling problem in the pattern recognition approach to machine translation. *Patt. Rec. Lett.*, 29(8):1145–1153, 2008.

[9] M. Pastor. *Aportaciones al reconocimiento automático de texto manuscrito*. PhD thesis, Univ. Politécnica de Valencia, October 2007.

[10] A. Pérez et al. Joining linguistic and statistical methods for Spanish-to-Basque speech translation. *Speech Comm.*, 50:1021–1033, 2008.

iAnaDoc books, journal articles & PhD thesis

[11] see [6].

[12] M. Ferrer. *Theory and Algorithms on the Median Graph. Application to Graph-based Classification and Clustering*. PhD thesis, Univ. Autònoma de Barcelona, 2008.

[13] J. Lladós. Advances in graphics recognition. In *Digital Document Processing: Major Directions and Recent Advances*, pages 285–304. 2007.

[14] J. Lladós and D. Blostein. *Special Issue on Graphics Recognition*, volume 9 of *IJDAR*. 2007.

[15] J. Lladós, W. Liu, and J.M. Ogier. *Graphics Recognition: Recent Advances and New Opportunities*, volume 5046 of *LNCS*. Springer Verlag, 2008. ISBN 978-3-540-88184-1.

[16] O. Ramos Terrades. *Linear Combination of multiresolution descriptors: applications to graphics recogntion*. PhD thesis, Universitat Autònoma de Barcelona, October 2006.

[17] O. Ramos Terrades, E. Valveny, and S. Tabbone. Optimal classifier fusion in a non-bayesian probabilistic framework. *IEEE PAMI*. (preprint).

[18] E. Valveny et al. A general framework for the evaluation of symbol recognition methods. *IJDAR*, 9(1):59–74, 2007.

iTransDoc conference papers

[19] V. Alabau et al. Improving speech-to-speech translation using word posterior probabilities. In *MT Summit XI*, 2007.

[20] V. Alabau et al. Inference of Stochastic Finite-State Transducers Using N-gram Mixtures. In *IbPRIA*. 2007.

[21] V. Alabau et al. Using posterior probabilities in lattice translation. In *IWSLT*, 2007.

[22] J. Andrés et al. Combining translation models in statistical machine translation. In *TMI*, 2007.

[23] J. Andrés and A. Juan. A phrase-based hidden markov model approach to MT. In *NAMT*, 2007.

[24] J.M. Benedí et al. Confidence measures for stochastic parsing. In *RANLP*, 2007.

[25] J.M. Benedí and J.A. Sánchez. Fast stochastic context-free parsing: a stochastic version of the valiant algorithm. In *IbPRIA*. 2007.

[26] J. Civera et al. Bilingual Text Cl. In *IbPRIA'07*.

[27] J. Civera and A. Juan. Domain adaptation in statistical machine translation with mixture modelling. In *WSMT*, 2007.

[28] J. Civera and A. Juan. Unigram-IBM Model 1 Mixtures for Bilingual Text Classification. In *LREC*, 2008.

[29] C. Estienne et al. Maximum entropy models for speech confidence estimation. In *ICASSP*, 2008.

[30] C. Estienne and A. Sanchis. A confidence measure for speech recognition systems based on two maximum entropy approaches. In *RPIC*, 2007.

[31] G. Gascó and J. A. Sánchez. Part-of-speech tagging based on mt techniques. In *IbPRIA*, 2007.

[32] G. Gascó and J.A. Sánchez. A* parsing with large vocabularies. In *RANLP*, 2007.

[33] A. Giménez and A. Juan. Bernoulli HMMs for Off-line Handwriting Recog. In *PRIS*, 2008.

[34] J. Gonazález and F. Casacuberta. Phrase-based finite state models. In *FSMNLP*, 2007.

[35] J. González and F. Casacuberta. A finite-state framework for log-linear models in Machine Translation. In *EAMT*, 2008.

[36] J. González and F. Casacuberta. Linguistic Categorisation in Machine Translation using Stochastic Finite State Transducers. In *Mixing Approaches to Machine Translation*, 2008.

[37] J. González et al. Learning finite state transducers using bilingual phrases. In *ICITP-CL'08*.

[38] J. González-Rubio et al. Translation applications under SisHiTra. In *L&T*, 2007.

[39] J. González-Rubio et al. A novel alignment model inspired on ibm model 1. In *EAMT*, 2008.

[40] J. González-Rubio et al. Optimization of log-linear machine translation model parameters using SVMs. In *PRIS*, 2008.

[41] S. Hasan et al. Triplet lexicon models for statistical machine translation. In *EMNLP*, 2008.

[42] J.Civera and A. Juan. Word alignment quality in the ibm 2 mixture model. In *PRIS*, 2008.

[43] M. Jha et al. Improved unsupervised speech recog. system using MLLR speaker adaptation and confidence measurement. In *VJTH*, 2008.

[44] A. Juan et al. Bridging the Gap between Naive Bayes and Maximum Entropy. In *PRIS*, 2007.
[45] A. Lagarda and F. Casacuberta. Applying boosting to statistical machine translation. In *EAMT*, 2008.
[46] A. Lagarda et al. Computer-assisted handwritten text transcription using speech recognition. In *VJTH*, 2008.
[47] M. Luján et al. A study on bilingual speech rec. involving a minority language. In *L&T'07*.
[48] M. Luján et al. El sistema de identificación de la lengua de prhlt. In *VJTH*, 2008.
[49] M. Luján et al. Evaluation of several maximum likelihood linear regression variants for language adaptation. In *LREC*, 2008.
[50] M. Luján et al. iATROS: A speech and handwritting recognition system. In *VJTH*, 2008.
[51] C. D. Martínez. On the training requirements for an automatic dialogue annotation technique. In *Discourse and Dialogue*, 2007.
[52] C.D. Martínez and V. Tamarit. Evaluation of different segmentation techniques for dialogue turns. In *LREC*, 2008.
[53] D. Ortiz et al. A general framework to deal with the scaling problem in phrase-based statistical machine translation. In *IbPRIA*. 2007.
[54] D. Ortiz et al. Phrase-level alignment generation using a smoothed loglinear phrase-based statistical alignment model. In *EAMT*, 2008.
[55] A. Pérez et al. Speech translation with phrase based stochastic finite-state transducers. In *ICASSP*, 2007.
[56] D. Pérez et al. The GERMANA database. In *ICDAR*, 2009. (submitted).
[57] L. Rodríguez et al. Computer assisted transcription of speech. In *IbPRIA*. 2007.
[58] V. Romero et al. Combination of N-grams and Stochastic Context-Free Grammars in an Offline Handwritten Recog. System. In *IbPRIA*. 2007.
[59] V. Romero et al. Computer Assisted Transcription for Ancient Text Images. In *ICIAR*. 2007.
[60] V. Romero et al. Explicit Modelling of Invariances in Bernoulli Mixtures for Binary Images. In *IbPRIA*. 2007.
[61] V. Romero et al. Aligning handwritten text images and transcriptions of historic documents. In *EVA*, 2008.
[62] V. Romero et al. Improvements in the computer assisted transcription system of handwritten text images. In *PRIS*, 2008.
[63] A. Sanchis et al. Estimation of confidence measures for MT. In *MT Summit XI*, 2007.
[64] G. Sanchis and F. Casacuberta. Reordering via N-best lists for Spanish-Basque translation. In *TMI*, 2007.
[65] G. Sanchis and F. Casacuberta. Increasing translation speed in phrase-based models via suboptimal segmentation. In *PRIS*, 2008.
[66] G. Sanchis et al. Improving interactive machine translation via mouse actions. In *EMNLP*, 2008.
[67] G. Sanchis et al. Introducing additional input inf. into IMT systems. In *JWMI-MLA*. 2008.
[68] G. Sanchis and J.A. Sánchez. Phrase segments obtained with SITGs for Spanish-Basque translation. In *VJTH*, 2008.
[69] G. Sanchis and J.A. Sánchez. Using parsed corpora for estimating stigs. In *LREC*, 2008.
[70] G. Sanchis and J.A. Sánchez. Vocabulary extension via POS information for SMT. In *Mixing Approaches to Machine Translation*, 2008.
[71] N. Serrano et al. GIDOC: Gimp-based Interactive transcription of old text DOCuments. In *ICDAR*, 2009. (submitted).
[72] V. Tamarit and C.D. Martínez. Dialog act labeling in the dihana corpus using prosody information. In *VJTH*, 2008.
[73] J. Tomás and F. Casacuberta. Phrase-based SMT using approx. matching. In *IbPRIA*. 2007.
[74] J. Tomás et al. Mining Wikipedia as a parallel and comparable corpus. In *CICLing*, 2008.
[75] A. H. Toselli et al. Viterbi based alignment between text images and their transcripts. In *LaTeCH*. 2007.
[76] A. H. Toselli et al. Computer assisted transcription of text images and multimodal interaction. In *JWMI-RMLA*. 2008.
[77] A.H. Toselli et al. Computer Assisted Transcription of Handwritten Text. In *ICDAR*. 2007.
[78] A. H. Tosellli et al. On-Line Handwriting Recognition System for Tamil Handwritten Characters. In *IbPRIA*. 2007.
[79] E. Vidal et al. Interactive pattern recognition. In *JWMI-RMLA*. 2007.

iAnaDoc conference papers

[80] A. Borràs and J. Lladós. Similarity-based object retrieval using appearance and geometric feature combination. In *IbPRIA*. 2007.
[81] A. Borràs and J. Lladós. Multi-scale layout descriptor based on Delaunay triangulation for image retrieval. In *VISAPP*, 2008.
[82] M. Delalandre et al. Building synthetic graphical documents for performance evaluation. In *GREC*, 2008.
[83] M. Delalandre et al. A fast cbir system of old ornamental letter. In *GREC)*, 2008.
[84] M. Delalandre et al. Performance evaluation of symbol recognition and spotting systems: An overview. In *DAS*, 2008.
[85] S. Escalera et al. Multi-class binary object categorization using blurred shape models. In *CIARP*, 2007.
[86] S. Escalera et al. Multi-class binary object categorization using blurred shape models. In *Progress in Pattern Recognition, Image Analysis and Applications*, LNCS, 2008.
[87] M. Ferrer et al. Bounding the size of the median graph. In *IbPRIA*, 2007.
[88] M. Ferrer et al. Comparison between two spectral-based methods for median graph computation. In *IbPRIA*, 2007.
[89] M. Ferrer et al. On the relation between the median graph and the maximum common subgraph of a set of graphs. In *GBRPR*, 2007.

[90] M. Ferrer et al. An approximate algorithm for median graph computation using graph embedding. In *ICPR*, 2008.

[91] M. Ferrer et al. Exact median graph computation via graph embedding. In *ISSPR*, 2008.

[92] M. Ferrer and E. Valveny. Combination of ocr engines for page segmentation based on performance evaluation. In *ICDAR*, 2007.

[93] A. Fornés et al. A dynamic time warping based method for classifying old handwritten musical symbols. In *Computer Vision: Advances in Research and Development*, 2007.

[94] A. Fornés et al. Handwritten symbol recognition by a boosted blurred shape model with error correction. In *IbPRIA*. 2007.

[95] A. Fornés et al. Old handwritten musical symbol classification by a dynamic time warping based method. In *GREC*, 2007.

[96] A. Fornés et al. Hand drawn symbol recognition by blurred shape model descriptor and a multiclass classifier. In *Graphics Recog.: Recent Advances and New Opportunities*. 2008.

[97] A. Fornés et al. Old handwritten musical symbol classification by a dynamic timewarping based method. In *Graphics Recognition: Recent Advances and New Opportunities*. 2008.

[98] A. Fornés et al. Writer identification in old handwritten music scores. In *DAS*, 2008.

[99] D. Karatzas. Detecting gradients in text images using the hough transform. In *DAS*, 2008.

[100] D. Karatzas et al. Segmentation robust to the vignette effect for machine vision systems. In *ICPR*, 2008.

[101] J. Lladós et al. Word spotting in archive documents using shape contexts. In *IbPRIA*, 2007.

[102] J. Lladós and J. Sánchez. Indexing historical doc. by word shape signatures. In *ICDAR*, 2007.

[103] J. Mas et al. An Incremental On-line Parsing Algorithm for Recognizing Sketching Diagrams. In *ICDAR*, 2007.

[104] J. Mas et al. Representing and parsing sketched symbols using adjacency grammars and a grid-directed parser. In *Graphics Recognition: Recent Advances and New Opportunities*. 2008.

[105] J. Mas et al. A semi-automatic annotation tool for archival documents. In *DAS*, 2008.

[106] O. Ramos et al. Optimal linear combination for two-class classifiers. In *ICAPR*, 2007.

[107] O. Ramos et al. A review of shape descriptors for document analysis. In *ICDAR*, 2007.

[108] J. A. Rodríguez et al. Advances in HMM-based handwritten Word spotting: features, statistical modelling and adaptation. In *Current challenges in Computer Vision*, 2007.

[109] J. A. Rodríguez et al. A pen-based interface for real-time document correction. In *ICDAR*, 2007.

[110] J. A. Rodríguez et al. Rejection strategies involving classifier combination for handwriting recognition. In *IbPRIA*, 2007.

[111] J. A. Rodríguez et al. Sketch-based document correction. In *Computer Vision: Advances in Research and Development*, 2007.

[112] J. A. Rodríguez et al. Categorization of digital ink elements using spectral features. In *Graphics Recog.: Recent Advances and New Opp.* 2008.

[113] J. A. Rodríguez et al. Unsup. writer style adapt. for handwritten word spotting. In *ICPR*, 2008.

[114] J. A. Rodríguez and F. Perronnin. Local gradient histogram features for word spotting in unconstrained handwritten doc. In *ICFHR*, 2008.

[115] J. A. Rodríguez and F. Perronnin. Score normalization for HMM-based word spotting using a universal background model. In *ICFHR*, 2008.

[116] P. P. Roy et al. A system to segment text and symbols from color maps. *LNCS*, 2007.

[117] P. P. Roy et al. Convex hull based approach for multi-oriented character recognition from graphical documents. In *ICPR*, 2008.

[118] P. P. Roy et al. Detection and recognition of multi-oriented text/symbols in graphical documents. In *Computer Vision: Progress of Research and Development*, 2008.

[119] P. P. Roy et al. Morphology based handwritten line segmentation using foreground and background information. In *ICFHR*, 2008.

[120] P. P. Roy et al. Multi-oriented english text line extraction using background and foreground information. In *DAS*, 2008.

[121] P. P. Roy et al. Recognition of multi-oriented touching characters in graphical documents. In *ICVGIP*, 2008.

[122] P. P. Roy, U. Pal, and J. Lladós. Multi-oriented character recognition from graphical documents. In *ICCR*, 2008.

[123] M. Rusinol et al. Boundary shape recognition using accumulated length and angle information. In *IbPRIA*. 2007.

[124] M. Rusinol et al. Camera-based graphical symbol detection. In *ICDAR*, 2007.

[125] M. Rusinol et al. A region-based hashing approach for symbol spotting in technical documents. In *Graphics Recognition: Recent Advances and New Opportunities*. 2008.

[126] M. Rusinol and J. Lladós. Graphical symbol spotting in technical documents using a region hash table. In *Computer Vision: Advances in Research and Development*, 2007.

[127] M. Rusinol and J. Lladós. Relational indexing of local descriptors to spot graphical objects in wiring diagrams. In *Current Challenges in Computer Vision*, 2008.

[128] M. Rusinol and J. Lladós. Word and symbol spotting using spatial organization of local descriptors. In *DAS*, 2008.

[129] E. Valveny et al. Performance characterization of shape descriptors for symbol representation. In *Graphics Recognition: Recent Advances and New Opportunities*. 2008.

[130] E. Valveny et al. Report on the third contest on symbol recognition. In *Graphics Recognition: Recent Advances and New Opportunities*. 2008.

[131] E. Valveny and M. Ferrer. Application of graph embedding to solve graph matching problems. In *CIFED*, 2008.