UNIVERSIDAD POLITÉCNICA DE VALENCIA
DEPARTAMENTO DE SISTEMAS INFORMÁTICOS Y COMPUTACIÓN

Interactive Transcription of Old Text Documents

Thesis
presented by Nicolás Serrano Martínez-Santos
supervised by
Dr. Alfons Juan Císcar
and
Dr. Jorge Civera Saiz

June 2, 2014

Interactive Transcription of Old Text Documents

Nicolás Serrano Martínez-Santos

Thesis performed under the supervision of doctors
Alfons Juan Císcar and Jorge Civera Saiz
and presented at the Universidad Politécnica de Valencia
in partial fulfilment of the requirements
for the degree Doctor en Informática

Valencia, June 2, 2014

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no 287755 (transLectures). Also supported by the EC (FEDER, FSE), the Spanish Government (MICINN, MITyC, "Plan E", under grants MIPRCV "Consolider Ingenio 2010", MITTRAL (TIN2009-14633-C03-01), iTrans2 (TIN2009-14511), and FPU (AP2007-02867), and the Generalitat Valenciana (grants Prometeo/2009/014 and GV/2010/067).

A mi tio Angelete,

Abstract

Nowadays, there are huge collections of handwritten text documents in libraries all over the world. The high demand for these resources has led to the creation of digital libraries in order to facilitate the preservation and provide electronic access to these documents. However text transcription of these documents images are not always available to allow users to quickly search information, or computers to process the information, search patterns or draw out statistics. The problem is that manual transcription of these documents is an expensive task from both economical and time viewpoints. This thesis presents a novel approach for efficient Computer Assisted Transcription (CAT) of handwritten text documents using state-of-the-art Handwriting Text Recognition (HTR) systems.

The objective of CAT approaches is to efficiently complete a transcription task through human-machine collaboration, as the effort required to generate a manual transcription is high, and automatically generated transcriptions from state-of-the-art systems still do not reach the accuracy required. This thesis is centered on a special application of CAT, that is, the transcription of old text document when the quantity of user effort available is limited, and thus, the entire document cannot be revised. In this approach, the objective is to generate the best possible transcription by means of the user effort available. This thesis provides a comprehensive view of the CAT process from feature extraction to user interaction.

First, a statistical approach to generalise interactive transcription is proposed. As its direct application is unfeasible, some assumptions are made to apply it to two different tasks. First, on the interactive transcription of handwritten text documents, and next, on the interactive detection of the document layout.

Next, the digitisation and annotation process of two real old text documents is described. This process was carried out because of the scarcity of similar resources and the need of annotated data to thoroughly test all the developed tools and techniques in this thesis. These two documents were carefully selected to represent the general difficulties that are encountered when dealing with HTR. Baseline results are presented on these two documents to settle down a benchmark with a standard HTR system. Finally, these annotated documents were made freely available to the community. It must be noted that, all the techniques and methods

developed in this thesis have been assessed on these two real old text documents.

Then, a CAT approach for HTR when user effort is limited is studied and extensively tested. The ultimate goal of applying CAT is achieved by putting together three processes. Given a recognised transcription from an HTR system. The first process consists in locating (possibly) incorrect words and employs the user effort available to supervise them (if necessary). As most words are not expected to be supervised due to the limited user effort available, only a few are selected to be revised. The system presents to the user a small subset of these words according to an estimation of their correctness, or to be more precise, according to their confidence level. Next, the second process starts once these low confidence words have been supervised. This process updates the recognition of the document taking user corrections into consideration, which improves the quality of those words that were not revised by the user. Finally, the last process adapts the system from the partially revised (and possibly not perfect) transcription obtained so far. In this adaptation, the system intelligently selects the correct words of the transcription. As results, the adapted system will better recognise future transcriptions. Transcription experiments using this CAT approach show that this approach is mostly effective when user effort is low.

The last contribution of this thesis is a method for balancing the final transcription quality and the supervision effort applied using our previously described CAT approach. In other words, this method allows the user to control the amount of errors in the transcriptions obtained from a CAT approach. The motivation of this method is to let users decide on the final quality of the desired documents, as partially erroneous transcriptions can be sufficient to convey the meaning, and the user effort required to transcribe them might be significantly lower when compared to obtaining a totally manual transcription. Consequently, the system estimates the minimum user effort required to reach the amount of error defined by the user. Error estimation is performed by computing separately the error produced by each recognised word, and thus, asking the user to only revise the ones in which most errors occur.

Additionally, an interactive prototype is presented, which integrates most of the interactive techniques presented in this thesis. This prototype has been developed to be used by palaeographic expert, who do not have any background in HTR technologies. After a slight fine tuning by a HTR expert, the prototype lets the transcribers to manually annotate the document or employ the CAT approach presented. All automatic operations, such as recognition, are performed in background, detaching the transcriber from the details of the system. The prototype was assessed by an expert transcriber and showed to be adequate and efficient for its purpose. The prototype is freely available under a GNU Public Licence (GPL).

Resumen

Actualmente existen grandes colecciones de documentos manuscritos en librerias de todo el mundo. La gran demanda de estos recursos ha llevado a la creación de librerías digitales para facilitar la preservación y el acceso electrónico a estos documentos. Sin embargo, la transcripción de las imágenes de estos documentos no está siempre disponible con tal de permitir la busqueda rápida y eficaz a los usuarios, o extraer patrones y datos estadísticos automáticamente. Esta tesis presenta una nueva aproximación para la transcripción asistida por ordenador (CAT) de documentos de texto manuscrito usando sistemas de reconocimiento de texto manuscrito (HTR).

El objetivo de las aproximaciones CAT es, completar de manera eficar una tarea de transcripción mediante la colaboración hombre-máquina, ya que el esfuerzo requerido para generar una transcripción manual es alto, y las transcriptiones obtenidas automáticamente por sistemas estado del arte aún no llegan a la precisión requerida. Esta tesis se centra en una aplicación especial de CAT, que es la transcripción de documentos manuscritos antiguos cuando el esfuerzo de usuario es limitado, y en consecuencia, el documento no puede ser revisado completamente. En esta aproximación, el objetivo es generar la mejor transcripción posible usando el esfuerzo de usuario disponible. Esta tesis ofrece una guia completa del proceso de CAT desde la extracción de características hasta la interacción de usuario.

Primero, se propone una aproximación estadística para generalizar la transcriptión interactiva. Dado que su aplicación directa es inabordable, se han realizado una serie de asunciones para aplicarla en dos tareas distintas: la transcripción interactiva de documentos de textos manuscritos y la detección del formato del documentos de texto.

A continuación, se describe el proceso de digitalización y anotación de dos documentos manuscritos antiguos reales. Este proceso se llevo a cabo dada la escasez de recursos similires y la necesidad de datos anotados con tal de comprobar todas las herramientas y técnicas desarrolladas en esta tesis. Estos dos documentos fueron escogidos cuidadosamente con tal de representar las típicas dificultades que se encuentran al emplear técnicas HTR. Se presentan resultados de referencia en estos dos documentos obtenidos con un sistema estándar para servir de referencia. Finalmente, estos documentos se han hecho públicos y accesibles

libremente a la comunidad. Hay que tener en cuenta que todas las técnicas y métodos desarrollados en esta tesis se han evaluado en estos dos documentos antiguos.

Seguidamente, se estudia y verifica de manera exhaustiva una aproximación CAT para HTR cuando el esfuerzo de usuario es limitado. El objetivo final de aplicar CAT se consigue mediante la unión de tres procesos separados. Dado el reconocimiento automático de un sistema HTR. El primer proceso consiste en localizar palabras (posiblemente) incorrectas y emplear el esfuerzo de usuario disponible en supervisarlas y corregirlas (si es necesario). Dado que la mayoría de las palabras no se van a supervisar ya que solo hay una cantidad limitada de esfuerzo de usuario, solo unas pocas serán seleccionadas para su supervisión. El sistema presenta al usuario un pequeño subconjunto de estas palabras elegidas por una estimación de su correctitud, o para ser más préciso, eligidas de acorde a su nivel de confianza. A continuación, el segundo proceso empieza una vez estas palabras de baja confianza han sido revisadas. Este proceso actualiza el reconocimiento del documento teniendo en cuenta las correcciones, lo cual mejora la calidad de las palabras que no han sido revisadas por el usuario. Finalmente, el último proceso adapta el sistema a partir de la última transcripción parcialmente supervisada (y posiblemente imperfecta) que se ha obtenido. En esta adaptación, el sistema escoge de manera inteligente que palabras correctas de la transcripción son usadas en la adaptación. Consecuentemente, el sistema adaptado reconocera mejor futuras transcripciones. Los experimentos de transcripción usando esta aproximación CAT que se han realizado muestran que esta aproximación es más eficaz cuando el esfuerzo de usuario aplicado es bajo.

La última contribución de esta tesis es un método para equilibrar la calidad de transcripción final y el esfuerzo de supervisión aplicado cuando se emplea la aproximación CAT previamente descrita. En otras palabras, este método permite al usuario controlar la cantidad de errores en las transcripciones obtenidas con una aproximación CAT. La motivación de este método es permitir a los usuarios decidir la calidad final deseada en los documentos, ya que una transcripción parcialmente erronea puede ser suficiente para entender el contenido, y el esfuerzo requerido para obtener esta transcripción puede ser significativamente menor que el de obtener una transcriptión manual completa. Consecuentemente, el sistema estima el esfuerzo de usuario mínimo requerido para alcanzar la cantidad de error definida por el usuario. La estimación del error se realiza calculando por separado el error causado por cada palabra reconocida, para después pedir al usuario que revisa aquellas donde hay más errores.

Además, se presenta un prototipo interactivo que integra la mayoria de las técnicas interactivas presentadas en esta tesis. Este prototipo se ha desarrollado para ser usado por expertos en paleografía, que no poseen ningún trasfondo en tecnologías HTR. Después de ser ajustado por experto en HTR, el prototipo permite a los transcriptores anotar un documento manualmente o utilizar la aproximación CAT presentada. Todos los procesos automáticos, como el reconocimiento, se ejecutan en segundo plano abstrayendo al transcriptor de los detalles internos del sistema. El prototipo fue probado por un experto transcriptor y demostró ser adecuado y eficiente para su finalidad. El prototipo está disponible libre y publicamente mediante una licencia GNU (GPL).

Resum

Actualment existeixen grans col·leccions de documents manuscrits en llibreries de tot el món. La gran demanda d'aquests recursos ha portat a la creació de llibreries digitals per tal de facilitar la preservació i access electrònic a aquests documents. No obstant, la transcripció de les images d'aquests documents no està sempre disponible per tal de permetre una cerca ràpida i eficaç als usuaris, o d'extraure patrons i dades estatístiques automàticament. Aquesta tesi presenta uno nova aproximació per a la transcripció asistida per ordinador (CAT) de documents de text manuscrit emprant sistems de reconeiximent de text manuscrit (HTR).

L'objectiu de les aproximacions CAT es, completar de manera eficaç una tasca de transcripció mitjançant la colaboració home-màquina, ja que l'esforç requerit per generar una transcripció manual es alt, i les transcripcions obtingudes automàticamente per sistemes estat del art encara no arriben a la precisió requerida. Aquesta tesi es centra en una aplicació especial de CAT, que es la transcripció de documents manuscrits antics quan l'esforç d'usuari es limitat, i en consequencia, el document no pot ser revisat completament. En aquesta aproximació, l'objectiu es generar la millor transcripció posible emprant l'esforç d'usuari disponible. Aquesta tesi ofereix una guia completa del proces de CAT desde l'extracció de característiques fins a l'interacció d'usuari.

Primer, es proposa una aproximació estadística per a generalitzar la transcripció interactiva. Donat que la seua aplicació directa es inabordable, s'han realitzat una serie d'assumcions per tal d'aplicar-la en dos tasques diferents: la transcripció interactiva de documents de texts manuscrits i la detecció del format de documents de text.

A continuació, es descriu el proces de digitalització i anotació de dos documents manuscrits antics reals. Aquest procés s'ha portat a terme donat el nombre escàs de recursos similars i la necessitat de dades anotades per tal de comprobar totes les ferramentes i tècniques desarrollades en aquesta tesi. Aquests dos documents han estat escollits amb cura amb l'objectiu de representar les típiques dificultats que es troben al utilitzar tècniques HTR. Es presenten resultat de referència en aquests dos documents obtinguts amb un sistema estàndar per tal de servir de referència. Finalment, aquests documents s'han fet públics i accessibles lliurement a la comunitat. Hi ha de tindres en compte que totes les tècniques i mètodes desarrollats en

aquesta tesí s'han evaluat en aquests dos documents antics.

Seguidament, s'estudia i verifica de manera exhaustiva una aproximació CAT per HTR quan l'esforç d'usuari es limitat. L'objectiu final d'aplicar CAT s'aconsegueix mitjançant l'unio de tres processos separats. Donat el reconeiximent automàtic d'un sistema HTR. El primer procés consisteix en localitzar paraules (possiblement) incorrectes i emprar l'esforç d'usuari disponible en supervisar-les i corregir-les (si es necessari). Donat que la majoria de les paraules no es van a supervisar ja que sols hi ha una quantitat limitada d'esforç d'usuari, sols unes poques seràn sel·leccionades per una estimació de la seua correctitut, o per a ser més precís, sel·lecionades d'acord amb el seu nivell de confiança. A continuació, el segon procés comença una vegada aquestes paraules de baixa confiança han estat revisades. Aquest procés actualitza el reconeiximent del document tenint en compte les correcions, el qual millora la qualitat de les paraules que no han estat revisades per l'usuari. Finalment, l'últim procés adapta el sistema a partir de l'última transcripció parcialment supervisada (i possiblement imperfecta) que s'ha obtés. En aquesta adaptació, el sistema escolleix de manera intel·ligent que paraules correctes de la transcripció son utilitzades en l'adaptació. En consequencia, el sistema adaptat reconeixerà millor les futures transcripcions. Els experiments de transcripció realitzats utilitzant aquesta aproximació CAT mostren que aquesta aproximació es més eficaç quan l'esforç d'usuari aplicat es baix.

L'última contribució d'aquesta tesi es un mètode per a equilibrar la qualitat de transcripció final i l'esforç de supervisió aplicat quan s'utilitza l'aproximació CAT previament descrita. En altres paraules, aquest mètode permeteix al usuari controlar la quantitat d'errors en les transcripcions obtèses amb una aproximació CAT. La motivació d'aquest mètode es permetre als usuaris decidir la qualitat final desitjada en els document, ja que una transcripció parcialment errònia pot ser sufficient per a entendre el contingut, i l'esforç requerit per obtindre aquesta transcripció pot ser significativament menor que el d'obtindre una transcripció manual completa. Com a resultat, el sistema estima l'esforç d'usuari mínim requerit per alcançar la quantitat d'error definit pel usuari. L'estimació del error es realitza calculant per separat l'error causat per cada paraula reconeguda, per a després demanar al usuari que revisé aquelles on hi ha més errors.

A més, es presenta un prototip interactiu que integra la majoria de les tècniques interactives presentades en aquesta tesi. Aquest prototip s'ha desarrollat per a ser utilitzat per experts paleogràfics, que no poseixen cap coneiximent de les tecnologies HTR. Després de ser ajustats per experts en HTR, el prototip permet als transcriptors anotar un document manualment o utilitzar l'aproximació CAT presentada. Tots els procesos automàtics, com el reconeiximent, s'executen en segón pla abstraent al transcriptor dels detalls interns del sistema. El prototip va ser probat per un expert transcriptor i desmostrà ser adecuat i eficient per a la seua finalitat. El prototip està disponible lliure i publicament mitjançant una llicencia GNU (GPL).

Preface

Nowadays, information of all types is stored on digital media, and can be almost instantly accessed by means of computer systems. However, until recently, information was stored in physical means in the form of handwritten scripts, and thus, there exists a considerable amount of handwritten old text documents in libraries all over the world. In the current digital era, electronic access to these documents is necessary in order to preserve handwritten documents and quickly accessing its contents. However, this task presents two main problems. First, a digital (scanned) version of a document is needed to preserve the original document. Next, experts are needed to transcribe the document, which is the most expensive and time consuming task of the whole process.

Natural Language Processing (NLP) is a research field that aims to develop computer systems able to automatically comprehend natural human language. HTR is an old but still hectic area of NLP, which deals with the transcription of handwritten text documents. The aim of HTR is to automatically generate the transcription of a given text image. Even though HTR has been studied for years, the quality of automatically transcribed documents is still unsatisfactory. These unsatisfactory results are caused in part by HTR systems, but an important factor is also the scarcity of annotated resources, from which these systems are estimated. A solution is to employ the benefits of automatic systems within the manual transcription of documents. These interactive solution is typically referred as CAT approach, in which the system is guided by a human, and the human is assisted by the system to complete the task as efficiently as possible.

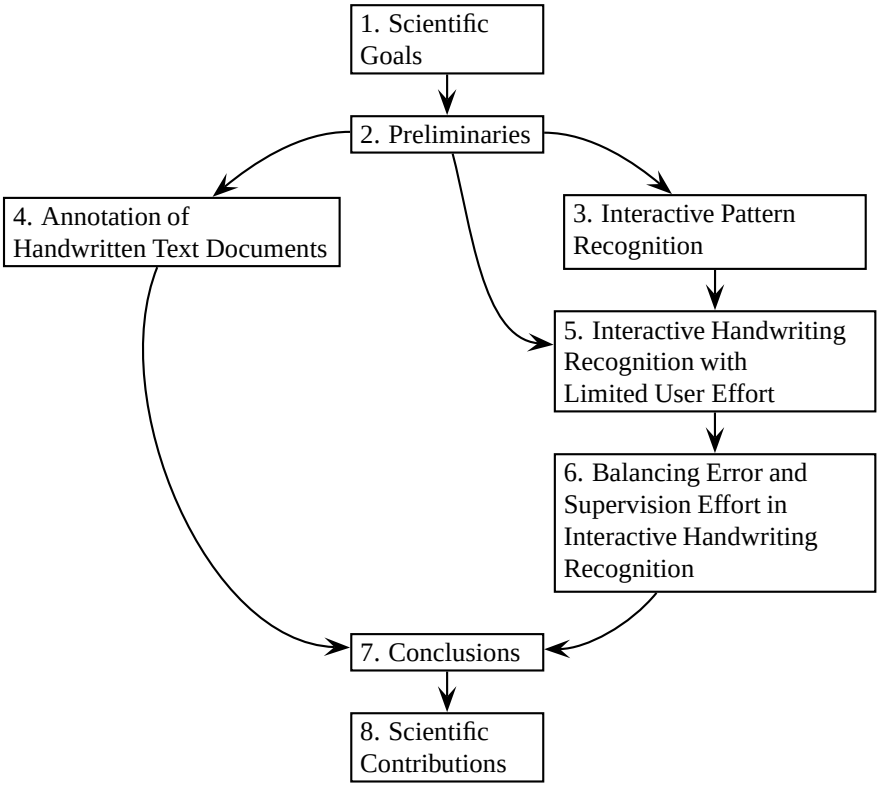
These approaches have been mainly focus on the efficient post-edition of system output, in which the user is asked to revise (and correct if necessary) certain parts of the system output. However, this effort could be employed in many multiple ways, and the interactive transcription proposed in this thesis presents alternative ways to the conventional output post-edition. A better approach is to fully exploit user effort by including his/her interactions in the transcription process, letting the system actively react to these interactions, improving the system performance. On the other hand, current CAT approaches deal with the transcription of the complete document, and even though user effort is saved when compared with the

manual transcription, it is not clear up to which extend. In fact, if the whole transcription of a document is required, the transcription has to be thoughtfully revised. Consequently, in cases in which the error rate is high, it may be better to complete the transcription manually.

However, there are applications in which a certain amount of errors may be tolerable. For instance, a limited quantity of user effort may be sufficient to generate an accurate enough transcription that conveys the meaning or useful for automatic search engines. Hence, the objective in this scenario to generate the best possible transcription given a certain amount of user effort. The solution to this problem is not straightforward, but can be approached using a sequential process of simple steps. The simplest way to employ the limited user effort is to correct erroneous words. However, the system needs to find these incorrect words to spare the user from this task. Additionally, these corrections reduce the uncertainty of the system, and thus, it can also be used to alter system decisions and improve its performance. Finally, even in case of perfect localisation of incorrect words, the user effort available could be insufficient to correct all of them. This causes that there might be errors in the resulting transcriptions. Effective adaptation from this partially supervised transcription could help to improve the system future performance.

The main objective of this thesis is to study and develop the proposed interactive approach, i.e. to interactively transcribe handwritten text documents when user effort is limited. This approach covers a wide range of techniques and algorithms, from the adaptation of HTR system and error estimation of a recognised transcription. Experimental results are presented on the transcription of two real handwritten text documents, which size is comparable to standard databases on HTR. It is worth noticing that even though the approach is applied to transcribe handwritten text document, it could also be used in other tasks implying the transcription of sequential data, such as speech or video.

The presented contributions are sequentially organised in 6 chapters that cover the work developed in this thesis. A sequential reading of the document is encouraged if the reader wish to learn about the complete work, but specific chapters can also be read attending to the following dependency graph:



First, Chapter 1 summarises the scientific goals of this work. Next, Chapter 2 introduces HTR, describing its history from its beginnings to the current state-of-the-art. Additionally, this chapter also explains the statistical foundations of HTR. The statistical foundations of the interactive pattern recognition proposed and its application to some problems in HTR is presented in Chapter 3. It includes tools for: document layout analysis, preprocessing, system training, line image recognition, and hypothesis verification. Two handwritten text databases are presented in Chapter 4, in which the digitisation and annotation process is thoroughly described. In this chapter, the description and validation of the baseline system that is used in the following chapters is also included.

Next, Chapter 5 describes the interactive transcription approach when user effort is limited, that has been developed in this thesis. It is based on the synergy employment of multiple techniques of different areas of Machine Learning (ML). First, active learning, which studies how to best improve a system from a limited number of new annotations, is used to locate incorrectly recognised words and ask the user to correct them, as they are expected to improve the system the most. Next, the current system hypothesis are update in a new Viterbi recomputation but constrained to newly user supervision, which help to the system to improve its recognition. Finally, techniques inspired on active and semisupervised learning are used altogether to adapt the current system, and thus improve its future performance. Methods to dynamically adjust the quantity of user effort applied to the interactive transcription are

described in Chapter 6. These methods estimate the error of the current recognition based on the previous system performance in order to obtain the user effort required for its correction.

Finally, Chapter 7 summarises the thesis contributions along with ideas for future work, while Chapter 8 sums up the scientific contributions of this work.

Contents

CHAPTER 1

Scientific Goals

In this chapter, we summarise the goals, which realisation resulted in the main contributions of this thesis.

Goals

The goals set up at the beginning of this work and that have been developed in this work are:

- Propose a complete interactive approach to transcribe handwritten old text documents.
- Create an interactive platform to enable users to interactively supervise any part of the HTR process.
- Study the application of an interactive transcription approach in cases in which user effort is limited.
- Take fully advantage of user interaction by only interacting those parts in the automatic transcription in which most benefit could be achieved.
- Create a system that automatically react to user interactions refining the resulting transcription.
- Study how to improve the system from its own output along with user interactions.
- Develop methods to calculate the degree of supervision needed, when the user decide on which error desires at then of the interactive process.
- Extract empirical results to assess the effectiveness of the proposed techniques and methods.

This thesis provides the solution to all this goals by studying and developing very different methods that collaborate in an interactive transcription platform. Concretely, the main contributions of this thesis are:

Implementation of an interactive transcription tool

CAT approaches need from users to complete efficiently the transcription task. The first contribution of this thesis is to develop an interactive prototype for transcribers. This prototype is a first step to detach expert paleographers from the details of HTR, enabling them to better transcribe text documents.

Annotation of two old text documents

HTR techniques need annotated documents in order to empirically demonstrate its correctness. However, nowadays, there are close to none old text documents that have been annotated. In this thesis, two old text documents have been digitalised, annotated and made freely available to the community.

Interactive HR with limited user effort

In order to efficiently employs a limited quantity of user effort in the transcription of a document a new CAT approach was created. This approach is divided in three processes. First, user effort is dedicated to supervise possibly incorrect words. Next, the transcription is improved from user corrections, updating the previous system hypothesis. Finally, the system is adapted from user supervised words and those unsupervised words that with a high probability are likely to be correct. This way, future transcriptions will be better.

Balancing error and user effort

Our final contribution is the creation of methods that estimate the effort required to obtain a transcription, using a CAT system, with a user defined amount of errors. Specifically, these methods calculate the error of the current system hypothesis. Then, the quantity of effort needed to reach the user requirements can be obtained.

CHAPTER 2

Preliminaries

Contents

2.1 Introduction

Natural language processing (NLP) is the research field that aims to develop computer systems able to automatically comprehend natural human language. NLP itself falls over two wider fields, Artificial Intelligence, and Computer Linguistics fields, as results, experts, methodologies, and theories from both fields converge to solve challenges produced by NLP. This thesis focus on an important area of NLP, Handwriting Text Recognition (HTR) in its application to the interactive transcription of old text documents when the user effort available is limited.

Handwritten Text Recognition (HTR) is an area of NLP, which deals with the transcription of handwritten text documents. The aim of HTR is to automatically generate the transcription of a given handwritten text image. The importance of HTR lies in the interest of libraries all over the world in transcribing their vast collections of documents in order to facilitate its access. Nowadays, this transcription task is carried at manually in an expensive and time-consuming task that can take up to 30 minutes per page (Pérez et al., 2009).

The first approaches to HTR were performed by means of tools and techniques from Optic Character Recognition (OCR), which can be considered solved even for hard scripts, such as Farsi (Liu et al., 2011; Mozaffari and Soltanizadeh, 2009). However, even though the HTR and OCR tasks seem similar, OCR systems are unable to deal with handwritten text documents. This is due to the difference between inputs for these tasks. On one hand, OCR deals with the transcription of a limited number of isolated characters in well-formed templates. Specifically, each character is recognised individually. On the other hand, HTR deals with unsegmented sequences of characters, drawn from non-uniform handwritten scripts. In this case, each character cannot be recognised by an OCR, as it cannot be correctly isolated. Nevertheless, HTR is highly related to Automatic Speech Recognition (ASR), as the two of them deal with the transcription of unsegmented signals, handwritten text images and speech, respectively. These similarities have caused that ASR techniques have been successfully applied to HTR (Bunke et al., 2004).

Even though ASR techniques have helped to improve the performance of HTR, and the HTR area has been studied for years, the quality of automatically transcribed documents is still unsatisfactory. One important reason is the complexity of the problem itself, as systems have to cope with several types of different handwriting styles. Another issue is the scarcity of annotated old text documents to train HTR systems. These problems have caused that to transcribe a given document, HTR systems have been converted into tools assisting the manual transcription process rather than fully automatic tools. The simplest approach consists in manually correcting the automatic transcription of an HTR system. However, this correcting process might be more time consuming than manually transcribing the document from scratch, as it requires the transcriber to revise the output of the system and correct it if necessary. This process can be more expensive than directly transcribing if there is a high quantity of errors in the system output. A better approach is to follow a computer assisted transcription (CAT) approach, in which the system is guided by a human, and the human is assisted by the system to complete the task as efficiently as possible. This CAT approach covers a wide range of techniques and tools, and thus, it can be approximated in many ways. For instance, a CAT system can be developed to complete the transcription task as efficiently as possible by asking the user to continuously correct transcription prefixes. This approach

has been implemented in HTR (Toselli et al., 2007) with encouraging results.

Most of CAT applications have been develop to assist the user in the entire transcription of a document. However, in these applications, despite the fact that user effort is reduced when compared to the manual transcription, the amount of effort required is unknown at the beginning. This causes that these well-studied applications cannot be applied when the quantity of supervision effort is limited. One reason for this limitation could be caused by of its cost, for instance transcription time or economic cost of a human transcriber. Additionally, human interaction is the bottleneck of the interactive approach, as the system has to wait before producing a result. Another important reason is that an error-free transcription might not be required. For instance, a partially erroneous transcription could be sufficient to convey the meaning, or it could be successfully used as input to search engines. In these cases, it is expected that the user effort needed to obtain this partially revised transcription is less than the complete manual case. Consequently, the objective of the approach is to obtain the best possible transcription by efficiently using the available user effort. This task involves many different steps. For example, error detection in order to ask the user only to correct the erroneous words, hence, saving user effort. The development and study of this approach, that to our knowledge have not been studied neither in HTR nor in related fields such as ASR, is the main topic of this thesis.

This chapter is organised as it follows. In the next section, we first review the current state-of-the-art in HTR. Afterwards, in Section 2.3 we briefly describe the steps that are performed to build a HTR system and then recognise a line image. Section 2.4 provides a brief description of theoretical details of HTR. Next, Section 2.5 describes the state-of-the-art of CAT approaches in HTR, and in Section 2.6, we give a brief explanation of our interactive transcription approach along with techniques and tools involved in its performance.

2.2 State-of-the-art in Handwritten Text Recognition (HTR)

In this section, a brief review of the history of HTR from its very beginnings to the current state-of-the-art systems is given. Previously, HTR and OCR system were described and clearly distinguished as they deal with different tasks. However, this distinction was not present at the beginning of their research, and as their development was very related, it might be confusing for the reader to clearly follow their progress. For the sake of clarity, in the following, OCR denotes the recognition of isolated (typewritten or handwritten) characters, while HTR refers to the recognition of (continuous and unsegmented) handwritten text.

HTR is reaching its maturity as its origins date back to the 50s with the application of the first OCR systems (Shepard, 1953). At that time, OCR could only handle typewritten characters from very restricted domains, such as certain fonts, Morse code or musical notes. Later, in the 60s, HTR systems were first applied for practical applications, such as transcription of postal codes or bank cheques. However, computer capacity those days could not handle large scale unconstrained domains, such as old text documents.

OCR techniques continued their development for two different inputs signals: online and offline. On one hand, online input signal stands for the one coming from the direct acquisition of pen movement derived of writing. Basically, online signal is composed by three dimensional vectors, which corresponds to the x-y coordinates for each time unit measured.

The first online systems were developed in the late 50s (Dimond, 1958), while the first commercial system appeared a few years later (Davis and Ellis, 1964). On the other hand, offline input corresponds to the acquisition of writing when it has been already written, that is typically extracted by scanning it from a physical document. Offline HTR is considered more difficult as the time correspondence of each pixel is lost, and recognition system have to relay in the writing order, for instance left-to-right in Latin script. OCR for both, online and offline, can be considered solved even for complex languages dealing with a high number of symbols (Liu et al., 2011; Mozaffari and Soltanizadeh, 2009). A detailed description of OCR from a historical point of view is described in (Mori et al., 1992).

As said, the HTR problem is to transcribe the contents of continuous handwritten text images. This problem is very similar to ASR, as the two areas study how to transcribe the corresponding words of an unsegmented input signal. In ASR case, the signal is composed by vectors of acoustic features for each time unit. Similarly, in HTR, the input signal is built from vectors of image features for each X coordinate unit in the image. ASR underlying techniques were first applied by Bunke et al. (1995) in HTR to transcribe isolated words in non-restricted domains. The main contribution of this approach was the use of Hidden Markov Models (HMMs) (Rabiner, 1990). HMMs are statistical models able to efficiently process unsegmented data. Later, HTR systems were leveraged by the inclusion of n-gram Language Models (LMs) to go from word to sentence (Bunke et al., 2004). This approach is still used in most of state-of-art HTR systems (Plötz and Fink, 2009).

Nowadays, HTR faces the problem of recognising an increasing number of different writing styles from any language. Script variability difficulties the generalisation needed in HTR systems. Additionally, language scarcity or complexity, difficulties the estimation of suitable LM, and even when it is possible, HTR systems must deal with large vocabularies. Current state-of-art approaches use additional steps and techniques over the basic approach to increment the system performance. For instance in (Dreuw et al., 2011), discriminative training is used to improve HMM estimation. Alternatively, in (España-Boquera et al., 2011), Neural Networks (NN) are used within the HMMs to improve their performance. Another successful approach uses recurrent NN (Graves et al., 2009). Despite the fact that great advances have been performed, HTR state-of-art systems only achieve recognition error rates around $[25\% - 35\%]$ in reference tasks, such as the IAM database (Marti and Bunke, 2002).

2.3 The Handwritten Text Recognition Process

In order to build an HTR system able to transcribe text line image we only need a set of annotated images. First of all, text line images have to be extracted. In order to complete this step, text line detection methods and Document Layout Analysis (DLA) have to be applied. However, this methods fall out this thesis and what is commonly known as HTR, thus, they will not be viewed in this section. Interested reader is referred to Appendix A for a brief description of these processes. In HTR, first, a *Preprocess* process is applied to the images in order to reduce the variability and noise within the images. Next, clean images are converted to numerical vectors better describing relevant features using a *Feature Extraction* method. Then, these feature vectors are used to build the HTR system in a phase called *Training*. Once the system has been trained, unannotated images can be transcribed in a phase called

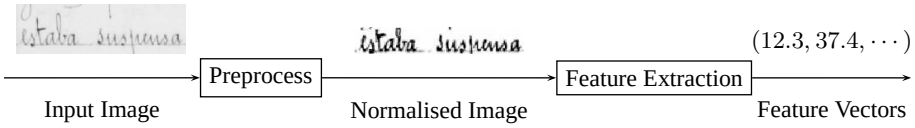


Figure 2.1: Preprocess and Feature Extraction phase in handwritten text recognition

Recognition.

The objective of the first two steps is to reduce the variability in text line images, and extract more informative features than pixel values, in order to improve the performance of the whole process. This phase is depicted in Figure 2.1. Preprocess is in charge of those techniques that modify the line image reducing its variability. This variability can be produced by many different factors such as noise or script slant. The results of the preprocess module is a cleaner image, in which letters are expected to share similar sizes, as observed in the example. On the other hand, the feature extraction step receives a clean image and transforms it into a vector of numerical features. These features are expected to better represent the most important characteristics within the image. This process can be motivated by expert decisions, such as Mel Feature Cesptral extraction in ASR (Young et al., 1995), or it can be performed by means of an automatic process that can transform input space into a more discriminative output space, such as Principal Components Analysis (PCA) (Jolliffe, 2002).

The second phase of the HTR process corresponds to the training of the models. As observed in Figure 2.2, the system takes a set of feature vectors and their corresponding transcription and it estimates a PR model to be used in the recognition phase. The internal theoretic details of this step are described in Section 2.4. The training of HTR models is a time-consuming task even when it is parallelled by grid computing. In fact, its cost has a linear dependency with the number of samples. Consequently, the more data is available for training, the better the recognition performance will be. However, it must be noted that, the performance gain from incrementing the available training data is not linear. In fact, some works have shown that, it is better not to use all the data available, but to intelligently select from which data to train (Hakkani-Tür et al., 2006).

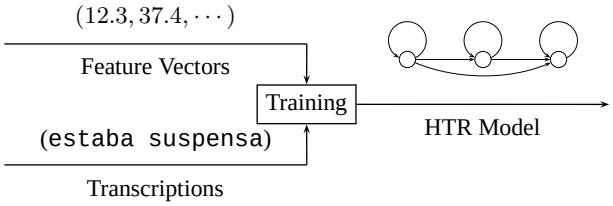


Figure 2.2: Training phase in handwriting text recognition

Due to the important computational cost required to train an HTR system, this step is typically performed offline. This cost also introduces an additional problem, when trying to learn from new annotated images become available. For this reason, system re-training is typically performed once a block of new annotated data is available. This fact is specially important in the work of this thesis, as re-training is a common step in the experimental setup

procedure. However, this limitation does not invalidate the results as in real applications training could be performed over night.

The last step, recognition, deals with the automatic transcription of unannotated images (in feature vector representation) using a HTR system, as it is depicted in Figure 2.3. Recognition is also a very time-consuming process because the best transcription is obtained by searching among all the possible hypothesis. However, this search can be efficient computed applying dynamic programming and pruning techniques. In fact, in current desktop computers, a text line image can be recognised every 30 seconds without performance degradation. However, this performance cannot still produce automatic transcription on real time base. So, similarly to the training phase, the recognition is typically performed offline, allowing the user to explore the transcription without waiting for the system.

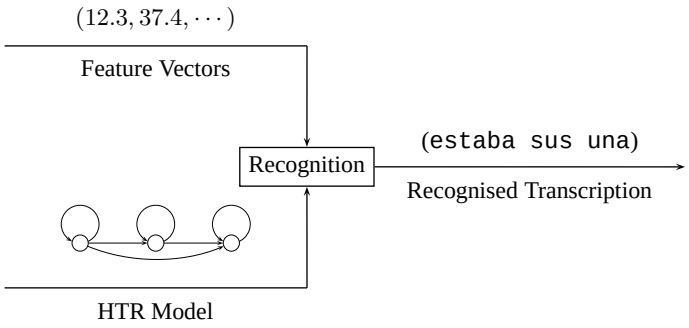


Figure 2.3: An overview of the handwriting recognition process

2.4 Theoretical Background of HTR

Current HTR systems are grounded on statistical PR techniques. PR is a subarea of Machine Learning (ML), which studies how to assign to a given input its corresponding label or class. In HTR, the input is defined as a sequence of T feature vectors $\mathbf{x} = x_1, \cdots, x_T$ representing the image, while the class label corresponds to a sequence of N words $\mathbf{w} = w_1, \cdots, w_N$ conforming the image. In the case of PR tasks in which the Classification Error Rate (CER) is used to measure the error, the best sequence of words $\mathbf{w}$, in terms of CER on the transcription, for the input $\mathbf{x}$ corresponds to the one maximising its posterior probability (Bishop, 2007)

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} p(\mathbf{w} \mid \mathbf{x}) \tag{2.1}$$

This posterior probability is factorised according to the Bayes rule as follows

$$\hat{\mathbf{w}} = \frac{\arg\max_{\mathbf{w}} p(\mathbf{x} \mid \mathbf{w}) p(\mathbf{w})}{p(\boldsymbol{x})} \tag{2.2}$$

where the term $p(\mathbf{x})$ remains constant for all the possible transcriptions and can be dropped in the maximisation. As result,

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} p(\mathbf{x} \mid \mathbf{w}) p(\mathbf{w}) \tag{2.3}$$

where $p(\mathbf{x} \mid \mathbf{w})$ is the probability density function describing how likely (or probable) is to observe $\mathbf{x}$ for the transcription $\mathbf{w}$, and $p(\mathbf{w})$ is the prior probability that expresses how likely is to observe the transcription $\mathbf{w}$.

As stated above, Bayes decision theory guarantees the optimal classification when the evaluation metric used is CER, and the probability distribution functions are known. However, these assumptions are not true in our case. First, the evaluation metric used in HTR is the Word Error Rate (WER), which is slightly different from CER (Schluter et al., 2011), therefore optimising the system in terms of CER may not improve its results in terms of WER. Last, probability distributions are unknown. In this work, we assume that there is no difference between the evaluations metrics, and that the probability distributions can be modelled statistically.

In this thesis, the conditional probability distribution $p(\mathbf{x} \mid \mathbf{w})$ is modelled using HMMs (Rabiner, 1990), and the prior distribution $p(\mathbf{w})$ is modelled using n-gram LMs (Chen, 1998).

2.4.1 Hidden Markov Character Models (HMMs)

PR typically deals with the classification of a given input into a single class. However, in many applications, the input may represent a structure or sequence of classes. For instance, in HTR, input is a text line image, and its classification is a sequence of words. The major problem in here is that the input is unsegmented, thus the alignment of which segment of the input generates which word in the transcription is unknown. This problem can be overcome using HMMs. HMMs have been successfully used since the 60s in fields such as, bioinformatics, ASR, or HTR. Their popularity is explained by their well defined mathematical properties, and their experimental good results.

As said in the previous section, we need to model the likelihood of a given sequence of feature vectors $\mathbf{x} = x_1, \cdots, x_T$ to be generated by the word transcription $\mathbf{w} = w_1, \cdots, w_N$, i.e. $p(\mathbf{x} \mid \mathbf{w})$. For the sake of simplicity, we consider the case of modelling the probability density function of a single word, so the latter probability $p(\mathbf{x} \mid \mathbf{w})$ will be expressed as probability as $p(\mathbf{x})$. Afterwards, in order to model the probability of a sentence, several word HMMs can be concatenated.

Direct estimation of $p(x)$ is unfeasible, as we should consider all possible segmentation of $\mathbf{x}$ into its corresponding transcription. To solve this problem, we assume that each element x_i of $\mathbf{x}$ has been produced (or emitted) in a different state q_i from a finite-state set $\mathcal{Q}$. As well as x_i elements q_i also follow a sequential order from 1 to T. A sequence of different states may represent a character or a word. We calculated the probability of $\mathbf{x}$ marginalising over $\mathbf{q}$

$$p(\mathbf{x}) = \sum_{\mathbf{q}} p(\mathbf{x}, \mathbf{q}) \tag{2.4}$$

in which the latter term can be expanded using the chain rule of probability

$$p(\mathbf{x}, \mathbf{q}) = \prod_{t=1}^{T} p(x_t, q_t \mid x_1^{t-1}, q_1^{t-1}) \tag{2.5}$$

We now make two further assumptions to approximate the last term. First, we assume that the probability of emitting x_t only depends on q_t. Last, we make a first order Markovian assumption in q_t, which implies that state q_t only depends on previous state q_{t-1}

$$p(x_t, q_t \mid x_1^{t-1}, q_1^{t-1}) = p(x_t, x_1^{t-1}, q_1^{t-1}) p(q_t \mid x_1^{t-1}, q_1^{t-1}) = p(x_t \mid q_t) p(q_t \mid q_{t-1}) \quad (2.6)$$

In Eq. 2.6, on one hand $p(x_t \mid q_t)$ corresponds to the emission probability, which is the probability of generating x_t on an state q_t. This emission probability could correspond to discrete tables, Gaussians, mixture of Gaussian, or Neural Networks. On the other hand, $p(q_i \mid q_{i-1})$ is the transition probability, which expresses the probability of *moving* from the state q_{i-1} to the state q_i.

HMMs are generative models, which model the emission of sequences of feature vectors $\mathbf{x}$. However, only the emitted sequence $\mathbf{x}$ is seen, while the sequence of states $\mathbf{q}$ remains *hidden*. This feature, in addition to the first order Markovian assumption is what gives origin to its name, *Hidden Markov*, to these models.

Given first order Markovian assumption, transitions from one state to the next only depends on the previous state. However, we need to define which states can be reached from a given one, along with their corresponding probability distributions $p(q_i \mid q_{i-1})$. This problem is solved by defining a stochastic finite-state automaton (Vidal et al., 2005), in which each state q_i corresponds to a state in $\mathcal{Q}$, and each edge represents a transition from state q_i to q_j

$$p(q_t = q_j | q_{t-1} = q_i) = a_{ij}, \forall_i \sum_j a_{ij} = 1 \quad (2.7)$$

where a_{ij} represents the transition probability, and thus the probability of the transitions going out from a state sum up to one

$$0 \leq a_{ij} \leq 1, \sum_j a_{ij} = 1 \quad (2.8)$$

The described automaton can be classified in different types according to its structure or topology. For instance, in an ergodic topology, every state of $\mathcal{Q}$ can be reached from any other. In our case, for sake of simplicity, and because of the sequential nature of $\mathbf{x}$ in HTR, we restrict the automaton to follow the so called left-to-right Bakis topology. In this topology, from one state q_i there are only three possible transitions. The loop transition going to the same state q_i, the next transition, which goes to the state q_{i+1}, and the skip transition, which goes to q_{i+2}. Figure 2.4 depicts an example of a three state Bakis topology.

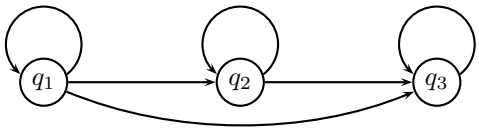


Figure 2.4: Example of Bakis topology

With the definitions we have made so far, we can now focus on two central issues of HMMs used in this thesis. First, given an HMMs and a set of training samples, we want to estimate its most likely parameters, which corresponds to the recognition step. Then, given an estimated HMM we want to compute the most likely sequence of states $\mathbf{q}$, which emits a

given $\mathbf{x}$. This task corresponds to finding the most likely alignment between a sequence states and a sequence of feature vectors, i.e. a text line image, and its corresponding transcription. This operation corresponds to the estimation of the first term in Eq. 2.3 in the training step.

The Learning Problem

The learning problem in HMMs is the problem of estimating the most likely parameters of an HMM, with a defined structure, given a set of training samples $\mathbf{X} = \{\mathbf{x_1}, \cdots, \mathbf{x_N}\}$, and their corresponding transcriptions $\mathbf{W} = \{\mathbf{w_1}, \cdots, \mathbf{w_N}\}$. In PR, the most likely parameters are typically obtained by means of maximum likelihood (ML) estimation (Duda et al., 2001)

$$L = \sum_{n=1}^{N} \log p(x_n \mid w_n) \tag{2.9}$$

ML estimation consist in obtaining the values of the parameters, which maximises the likelihood of the training samples, assuming that they are i.i.d. (independent and identically distributed random variables). In the case of HMMs, direct maximisation of this function leads to a complex equation, in which there is no closed form for the maximisation. However, as introduced earlier, the probability in Eq. 2.6, can be decomposed with a latent variable $\mathbf{q}$, defining a new model, with unobserved latent variables.

The estimation of this new model can be carried out by the Expectation-Maximisation (EM) algorithm (Dempster et al., 1977). The EM algorithm proposes to maximise the expected ML given the latent variable. Dempster et al. (1977) showed that a local optimum on this function corresponds to a local optimum in the ML function, and thus to a valid estimation of the parameters. EM maximisation implies two different steps: the Expectation (E) step, and the Maximisation (M). EM algorithm starts with an initial value of the model parameters. As the algorithm is demonstrated to guarantee convergence to a local optimum, the initialisation can be performed randomly. However, different initialisation can lead to different local optimum. In this thesis, we calculated the initial HMM parameters by uniformly splitting the training samples to each visual character HMM and each of their states, estimating its mean values. This initialisation is based on the standard initialisation method employed by the known HTK toolkit (Young et al., 1995).

In the E step, the current model parameters are used to find the posterior distribution of the latent variables, and their corresponding expected values. In HMMs, the E step can be performed in two different ways. First, we can follow the E step definition directly and estimate the expected values itself by means of the Forward-Backward (or Baum-Welch) algorithm (Bishop, 2007), or we can perform a maximisation as an approximation to the expected value (Neal and Hinton, 1998), which is obtained by the algorithm presented in next section.

In the M step, we maximise the model parameters according to the newly estimated expected latent variables. In HMMs, the expected value for the latent variables can be considered as weighted paths emitting each sample. In consequence, these paths aligns feature vectors $\mathbf{x}$ with each state of the HMM, which model parameters can be directly estimated by means of ML. In our case, there are two distinct set of model parameters: the transition probabilities and the emission probabilities. The transition probabilities are directly estimated as

the expected values accounts each time the transition was used over the rest. On the other hand, emission probabilities are estimated according to their underlying model. In this thesis, these probabilities are modelled using Gaussian Mixture Models (GMM) (Duda et al., 2001). So, a gaussian mixture model is trained in each state from the $\mathbf{x}$ that were aligned in the E step by applying ML estimation.

The Decoding Problem

The decoding problem in HMMs is the problem of finding the most probable state sequence $\mathbf{q}$, which generates a given input sample $\mathbf{x}$. The most probable state sequence $\hat{\mathbf{q}}$ is calculated as

$$\hat{\mathbf{q}} = \arg\max_{\mathbf{q}} \prod_{t=1}^{T} p(x_t \mid q_t) p(q_t \mid q_{t-1}) \tag{2.10}$$

which can be recursively extracted

$$\hat{\mathbf{q}} = \arg\max_{q_1} \left\{ \arg\max_{q_2} \left\{ \cdots \arg\max_{q_T} \left\{ p(x_T \mid q_T) p(q_T \mid q_{T-1}) \right\} \cdots \right\} \right\} \tag{2.11}$$

We define the Viterbi recursion (Viterbi, 1967) function of a state j as it follows

$$v_t(j) = \arg\max_{i, i \in \mathcal{Q}} v_{t-1}(i) p(x_t \mid q_j) p(q_j \mid q_i) \tag{2.12}$$

which is efficiently computed using the Viterbi algorithm (Forney, 1973)

$$v_t(j) = \begin{cases} a_{0j} b_{j1} & t = 1 \quad (2.13) \\ \left\{ \max_i v_{t-1}(i) \right\} a_{ij} b_{jt} & \text{otherwise} \quad (2.14) \end{cases}$$

where a_{ij} corresponds to the transition probability $p(q_i \mid q_j)$ and b_{jt} corresponds to the emission probability $p(x_t \mid q_j)$.

The Viterbi algorithm is a case of dynamic programming, in which we compute from x_1 to x_T and for each $q_j \in \mathcal{Q}$ its corresponding $v_t(j)$. In the end, the state that maximises the function v_T is obtained, which allows to find its most probable predecessor, extracting the most probable path $\hat{\mathbf{q}}$.

2.4.2 n-gram Language Models

LM estimation deals with the task of modelling the probability of a given sentence $\mathbf{w} = \{w_1, \cdots, w_N\}$. This is a core task in NLP tasks such as HTR, as it directly corresponds to the second part of the classification equation in Eq. 2.3. LM has been studied for two decades (Rosenfeld, 2000), and a wide variety of different models have been developed. One of the most successful and used models are the n-gram models (Goodman, 2001). Given a sentence $\mathbf{w}$, we decompose its probability by means of the chain rule

$$p(\mathbf{w}) = \prod_{t=1}^{T} p(w_t \mid w_1^{t-1}) \tag{2.15}$$

where $p(w_t \mid w_1^{t-1} =$ is the probability of observing w_t once w_1^{t-1} has occurred.

We could directly estimate each term of the product, however, the number of parameters exponentially grows with the length of $\mathbf{w}$. Therefore, we make a Markovian assumption of order n, which means that each word only depend on the preceding $(n-1)$ words

$$p(\mathbf{w}) \approx \prod_{t=1}^{T} p(w_t \mid w_{t-(n-1)}^{t-1}) \tag{2.16}$$

We can now limit the number of parameters by choosing a suitable n. For instance, given a text with W different words, an n-gram would have at most W^n parameters. It must be noted that, the first terms of the previous equation do not posses the needed history to be correctly estimated. This problem is solved by adding $n-1$ times at the beginning of the sentence the special word "<s>". This way the probability of word to occur at the beginning can be calculated.

Given a text of W words, we want to estimate each of the n-gram model probabilities

$$p(w \mid h) \; \forall w \in W, \forall h \in W^{n-1} \tag{2.17}$$

in which w corresponds to each word of the lexicon, and h to each possible history h of length $n-1$. N-gram parameters can be estimated by ML estimation as

$$p(w \mid h) = \frac{N(h,w)}{N(h)} \tag{2.18}$$

in which the N function accounts for the number of times a certain event has been observed. However, available data is usually scarce to estimate the large number of parameters even for small values of n. This is mainly caused because word n-gram events in natural languages follow the Zipf's law (Zipf, 1949). Zipf's law states that an event frequency is proportional to its rank in the frequency table. For instance, the most frequent event will appear almost twice as ofter as the second, three times more than third, and so on. Thus, the quantity of text needed to effectively, correctly estimate the n-gram parameters is far unattainable. To solve this problem, smoothing techniques are used.

Smoothing techniques are based on the idea of discounting probability mass from observed events, and its redistribution into unobserved events. In this thesis, we only describe the modified Knesser-Ney smoothing (Chen, 1998), as it is the one that performed best when selecting the optimum smoothing technique for the thesis experiments. In this smoothing, the probability of each n-gram is estimated considering all its corresponding lower order n-grams

$$p(w_i \mid w_{i-(n-1)}^{i-1}) = \alpha(w_i \mid w_{i-(n-1)}^{i-1}) + \gamma(w_{i-(n-1)}^{i-1})\alpha(w_i \mid w_{i-(n-2)}^{i-1}) \tag{2.19}$$

The α function is estimated by a slightly modification of the Knesser-Ney smoothing method (Kneser and Ney, 1995), in which each n-gram is discounted a quantity according to the number of times it has occurred. The γ is a scale factor so the probability sums up to one.

LMs are evaluated in terms of perplexity. Perplexity is a quality measurement in information theory. Given a discrete probability distribution p, which in our case corresponds to

an n-gram model, and a set of T sentences $\mathbf{W} = \{\mathbf{w}_1, \cdots, \mathbf{w}_N\}$, the perplexity (PP) is calculated as

$$PP(\mathbf{W}) = 2^{-\frac{1}{N}\sum_{t=1}^{T}\log_2 p(\mathbf{w}_t)} \tag{2.20}$$

where N corresponds to the number of words in T. The perplexity corresponds exactly to two power of the entropy of $\mathbf{W}$ given our n-gram LM, which can be interpreted as the expected number of words after a given one. LM with less perplexity are better estimated as the uncertainty of which words follows after is smaller.

2.5 Interactive HTR

Even though there exists many automatic systems dealing with different tasks with high performance, there also exists a high number of problems, such as transcription of text document, in which a fully automatic approach is unfeasible. As said, manual transcription of old text documents is very expensive in economic and time terms, and unfortunately, text documents cannot be transcribed with acceptable results by current state-of-the-art HTR system. However, as in other daily tasks, a synergy can be achieved by combining the best of both, human and machine, in a CAT approach. In this approach, the system and the user help each other in order to efficiently complete the task, that is minimising user effort. This user effort can be provided at different levels. In HTR, for instance, España-Boquera et al. (2011) employed user effort to train a NN to reduce the variability of input images. On the other hand, Agua et al. (2012) showed the transcription of multilingual documents can be improved by manually specifying the language i which each line is written. However, the most common use of user interaction is to supervise the system output, in order to obtain a correct transcription.

The application of CAT approaches altogether with PR techniques is not new. In fact, they have been used in a wide range of different areas, such as bioinformatics (Doi, 2007) or ASR. In ASR for instance (Barras et al., 2001), a first step was to automatically recognise an audio segment, and then, manually correct it with an interactive tool that enables the user to efficiently navigate through speech. Specifically, the objective of these tools is to facilitate the transcription to the user, when compared with the tedious manual transcription. Other more refined approaches in ASR locate errors and pass them to the user (Luz et al., 2008), further reducing the effort (Hakkani-Tür et al., 2006). CAT approaches are not new either in the transcription of old text documents. In the DEBORA project (Bourgeois and Emptoz, 2007), an approximation for CAT in the OCR of old machine printed documents was presented. In this approach, the user attention is based on in correcting those system transcriptions that could not be automatically classified. Similarly, reCAPTCHA (Ahn et al., 2008) employs user correction to transcribe difficult printed documents while also serving as additional protection when filling web forms.

Hitherto we have introduced some basic CAT approaches when user effort is employed in correcting the system output. The basic features of these approaches are the use of adequate interactive tools to navigate the image, together with an error detection tool, that highlights possibly incorrect words to the user . However, if errors occur frequently, it could be better to ignore the system output and complete the task manually (H.Nanjo and T.Kawahara, 2006; Luz et al., 2008). A bettter idea is to employ user interaction beyond the simple correction of the system output. For instance, in HTR, an incorrectly recognised word of a given text line,

typically affects the surroundings words, generating more errors. When the user supervises a recognised word, the uncertainty of the system around that word is reduced, and thus the transcription may improve (Culotta et al., 2006).

In this regard, one of the most successful CAT approaches is the prefix-based approach. The main idea of this approach is to improve the current system hypothesis by recomputing it constrained to a correct prefix. Concretely, first, the user validates the prefix of a system hypothesis up to the first incorrect word, which is corrected. Next, the validated prefix and the user corrected word are employed to predict the remaining suffix by constraining the search process. This process is repeated until the whole transcription has been revised. This approach has been the base of many works dealing with very different applications, such as HTR (Toselli et al., 2007), ASR (Revuelta-Martínez et al., 2012) or syntactic tree annotation (Sánchez-Sáez et al., 2010). All these approaches successfully reduce the effort needed to obtain the required output.

2.6 Interactive HTR in this thesis

In the problem that is studied in this thesis, the interactive HTR with limited user effort, the previously presented approaches present a major drawback. In fact, even though when using this approaches, the user effort required for transcribing the document is lower than in case of manual transcription, it is not easy to estimate how much is required. It would be better, for applications in which the effort is limited, to take the most advantage of the available user effort and produce the highest quality transcription possible. In other words, the objective of this new approach is, given a quantity of user effort, obtaining the best transcription possible.

The most straightforward way to develop the newly presented approach is to invest the limited quantity of user effort in supervising only those recognised parts which have been incorrectly recognised. The first step is to decide at which level the supervision is going to be applied. In some works of ASR (Hakkani-Tür et al., 2006), this supervision was performed at the sentence level, because, it may be difficult or unnatural to the user to correct isolated audio segments. For instance, when supervising sentences, the user effort would be employed to correct the most erroneous ones. However, a better approach would be to only supervise the incorrect words within those sentences. In fact, in HTR, words can be isolated, and presented to the user in closed boxes, as in the successful reCAPTCHA (Ahn et al., 2008).

In order to select possibly incorrect words, an additional step is added to the process, in which the HTR recognition system scores its output according to its reliability with the current hypothesis. When these scores manage to discriminate which words are correct or incorrect, they can be used as *Confidence Measures* (CM), i.e. an score of the system uncertainty on a given word. In consequence, words with low CMs would then correspond to possibly mis-recognised system hypothesis. Examples of valid CMs are system scores, such as the likelihood, or external features, such as morphological classification, or a combination of them. CMs have been studied and applied in a wide range of areas, such as machine translation (Ueffing et al., 2003), or ASR (Wessel et al., 2001). In this thesis, we have mainly used word posterior probabilities from Eq. 2.1, as CMs in two different applications. First, to select low confidence recognised words that will be supervised by the user. Second, to select high confident words to improve system via adaptation.

CMs enable us to detect incorrect words, however, we need a method to select which of these words will improve the most the transcription quality. Given a set of recognised samples along with their CMs, and a limited quantity of supervision effort, *Active Learning* (AL) (Settles, 2010) is a research area that studies how to efficiently use this user effort in order to improve the most the current system. It must be noted that, this application is closely related to CAT, however, in CAT the finality is to select which words better improves the final transcription, not the system performance. AL have been successfully employed in different areas and applications. For instance, in applications where annotation is very expensive, AL helps to select a small set of samples, which obtain a system with an acceptable performance. Another application is effective adaptation from few samples. One of the most widespread and successful AL techniques is uncertainty sampling, which selects the recognised samples to be annotated according to their confidence. Low confidence words will be likely produced by poorly estimated models or unknown events, thus its correction will include them into the training set, improving the system performance.

Once the user has supervised (and corrected if necessary) those recognised words selected by the system, we obtain a partially supervised transcription. Supervised words can be directly used to train our system, as they correspond to valid (annotated) samples. Nevertheless, unsupervised words may have been correctly recognised, and they could be included into the training set. The first approach is to include all unsupervised words in the training, which is called unsupervised learning. Unfortunately, as it has been shown in previous works, the improvement is quite limited (Serrano et al., 2009). However, as AL techniques select for supervision low confidence words, the rest unsupervised samples should correspond to high confidence ones. As a result, an effective selection of unsupervised words would help to improve system performance, or at least, to remain unaltered. In fact, *Semisupervise Learning* (SL) studies this problem (Zhu, 2006). SL has been applied in ASR (Wessel and Ney, 2005), and HTR (Frinken et al., 2011), as it reduces the amount of annotated samples needed in a task. The most simple yet effective technique in SL is called self-training, which uses CMs to select which words are used to adapt the system. High scored CMs represent high confidence words, which are likely to be correct.

User interaction is a useful resource. When dealing with the supervision of a word or a sentence, its supervision may help to improve the words before and after that being corrected, as there is a direct dependency between them. This approach could help to improve a transcription after the user has supervised a few recognised words. For instance, the supervision of a word may include a new word into the system vocabulary, or increment the confidence of the surrounding words. This latter approach has been followed in a wide range of areas, such as information retrieval (Kristjannson et al., 2004), or HTR. A successful but limited approach was presented by Toselli et al. (2007). They presented a technique that given a text line image, the user supervises the its prefix, correcting the first incorrect word that is found. Next, the system generates the most probable suffix constrained to this correct prefix. They shown that this prefix constraining help to improve the suffix, and consequently, the quality on final transcriptions. However, in this thesis, as any word within the line can be supervised, a *Constrained Viterbi search* algorithm has been develop to recognise samples in which some words have been supervised. This words will reduce the number of hypothesis to be considered, and therefore, a better transcription will be obtained.

In this thesis, we have integrated and extended all the previously described techniques

and tools: *Confidence Measures*, *Active Learning*, *Semisupervise Learning* and *Constrained Viterbi search*, into the a CAT approach to transcribe old text documents.

Bibliography

M. Agua, N. Serrano, J. Civera, and A. Juan. Character-based handwritten text recognition of multilingual documents. In *Proc. of Advances in Speech and Language Technologies for Iberian Languages (IBERSPEECH 2012)*, pages 187–196, 2012.

L. V. Ahn, B. Maurer, C. Mcmillen, D. Abraham, and M. Blum. reCAPTCHA: Human-based character recognition via web security measures. *Science*, 321: 1465–1468, 2008.

C. Barras, E. Geoffrois, Z. Wu, and M. Liberman. Transcriber: Development and use of a tool for assisting speech corpora production. *Speech Communication*, 33(1-2):5 – 22, 2001.

C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 1st ed. 2006. corr. 2nd printing edition, 2007.

F. L. Bourgeois and H. Emptoz. DEBORA: Digital AccEss to BOoks of the RenAissance. *Int. Journal on Document Analysis and Recognition (IJDAR)*, 9:193–221, 2007.

H. Bunke, M. Roth, and E. Schukat-Talamazzini. Off-line cursive handwriting recognition using hidden markov models. *Pattern Recognition*, 28(9):1399 – 1413, 1995.

H. Bunke, S. Bengio, and A. Vinciarelli. Offline recognition of unconstrained handwritten texts using HMMs and statistical language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(6):709 –720, 2004.

S. F. Chen. An empirical study of smoothing techniques for language modeling. Technical report, 1998.

A. Culotta, T. Kristjansson, A. McCallum, and P. Viola. Corrective feedback and persistent learning for information extraction. *Artificial Intelligence*, 170(14):1101–1122, 2006.

M. R. Davis and T. O. Ellis. The RAND tablet: a man-machine graphical communication device. In *Proc. of the 1964 fall joint computer conference, part I (AFIPS 1964)*, pages 325–331, 1964.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1): 1–38, 1977.

T. L. Dimond. Devices for reading handwritten characters. In *Eastern joint Computer Conference (IRE-ACM-AIEE 1957)*, pages 232–237, 1958.

K. Doi. Computer-aided diagnosis in medical imaging: Historical review, current status and future potential. *Computerized Medical Imaging and Graphics*, pages 198–211, 2007.

P. Dreuw, G. Heigold, and H. Ney. Confidence and Margin-Based MMI/MPE Discriminative Training for Offline Handwriting Recognition. *International Journal on Document Analysis and Recognition (IJDAR)*, 14(3):273–288, 2011.

R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification (2nd Edition)*. Wiley-Interscience, 2001.

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

J. Forney, G.D. The Viterbi algorithm. *Proc. of the IEEE*, 61(3):268 – 278, 1973.

V. Frinken, A. Fischer, H. Bunke, and A. Foornes. Co-training for handwritten word recognition. In *Proceedings of the 11th International Conference on Document Analysis and Recognition (ICDAR 2011)*, pages 314–318, 2011.

J. T. Goodman. A bit of progress in language modeling. Technical report, 2001.

A. Graves, M. Liwicki, S. Fernández, R. Bertolami, H. Bunke, and J. Schmidhuber. A novel connectionist system for unconstrained handwriting recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(5):855–868, 2009.

D. Hakkani-Tür, G. Riccardi, and G. Tur. An active approach to spoken language processing. *ACM Transactions on Speech and Language Processing*, 3:1–31, 2006.

Y. H.Nanjo and T.Kawahara. Computer assisted speech transcription system for efficient speech archive. In *Proc. of the 2006 Western Pacific Acoustics Conference (WESPAC 2006)*, 2006.

I. Jolliffe. *Principal Component Analysis*. Springer Series in Statistics. Springer, 2002.

R. Kneser and H. Ney. Improved backing-off for m-gram language modeling. In *Proc. of the 30th Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 1995)*, pages 181–184, 1995.

T. Kristjannson, A. Culotta, P. Viola, and A. McCallum. Interactive information extraction with constrained conditional random fields. In *Proc. of the 19th National Conference on Artificial Intelligence (AAAI 2004)*, pages 412–418, 2004.

C.-L. Liu, F. Yin, Q.-F. Wang, and D.-H. Wang. Icdar 2011 chinese handwriting recognition competition. In *Proc. of the 11th International Conference on Document Analysis and Recognition (ICDAR 2011)*, pages 1464 –1469, 2011.

S. Luz, M. Masoodian, and B. Rogers. Interactive visualisation techniques for dynamic speech transcription, correction and training. In *Proc. of the 9th Int. Conf. on Human-Computer Interaction (CHINZ 2008)*, pages 9–16, Wellington, New Zealand, 2008.

U. V. Marti and H. Bunke. The IAM-database: an English sentence database for off-line handwriting recognition. *International Journal on Document Analysis and Recognition (IJDAR)*, pages 39–46, 2002.

S. Mori, C. Suen, and K. Yamamoto. Historical review of ocr research and development. *Proc. of the IEEE*, 80(7):1029 – 1058, 1992.

S. Mozaffari and H. Soltanizadeh. Icdar 2009 handwritten farsi/arabic character recognition competition. In *Proc. of the 10th Int. Conf. on Document Analysis and Recognition (ICDAR 2009)*, pages 1413 –1417, july 2009.

R. Neal and G. Hinton. A View of the EM Algorithm that Justifies Incremental, Sparse, and other Variants. In M. I. Jordan, editor, *Learning in Graphical Models*. Kluwer, 1998.

D. Pérez, L. Tarazón, N. Serrano, O. Ramos-Terrades, and A. Juan. The GERMANA database. In *Proc. of the 10th Int. Conf. on Document Analysis and Recognition (ICDAR 2009)*, pages 301–305, Barcelona (Spain), 2009.

T. Plötz and G. A. Fink. Markov models for offline handwriting recognition: a sur-

vey. *Int. Journal of Document Analysis and Recognition (IJDAR)*, 12(4):269–298, 2009.

L. R. Rabiner. In *Readings in speech recognition*, chapter A tutorial on hidden Markov models and selected applications in speech recognition, pages 267–296. 1990.

A. Revuelta-Martínez, L. Rodríguez, and I. García-Varea. A computer assisted speech transcription system. In *Proc. of the 13th Conf. of the European Chapter of the Association for computational Linguistics (EACL 2012)*, pages 41–45, 2012.

R. Rosenfeld. Two decades of statistical language modeling: Where do we go from here. In *Proc. of the IEEE*, pages 1270–1278, 2000.

R. Sánchez-Sáez, L. A. Leiva, J.-A. Sánchez, and J.-M. Benedí. Interactive predictive parsing using a web-based architecture. In *Proc. of The 11th Annual Conf. of the North American Chapter of the Association for Computational Linguistics Demonstration Session (HLT-DEMO 2010)*, pages 37–40, Stroudsburg, PA, USA, 2010.

R. Schluter, M. Nussbaum-Thom, and H. Ney. On the relationship between bayes risk and word error rate in asr. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(5):1103 –1112, july 2011.

N. Serrano, D. Pérez, A. Sanchis, and A. Juan. Adaptation from partially supervised handwritten text transcriptions. In *Proc. of the 11th Int. Conf. on Multimodal Interfaces and the 6th Workshop on Machine Learning for Multimodal Interaction (ICMI-MLMI 2009)*, pages 289–292, 2009.

B. Settles. Active learning literature survey. Technical report, 2010.

D. Shepard. Apparatus for reading, 12 1953.

A. Toselli, V. Romero, L. Rodríguez, and E. Vidal. Computer assisted transcription of handwritten text. In *Proc. of the 9th Int. Conf. on Document Analysis and Recognition (ICDAR 2007)*, pages 944–948, Curitiba (Brazil), 2007.

N. Ueffing, K. Macherey, and H. Ney. Confidence measures for statistical machine translation. In *In Proc. of the 9th Machine Translation Summit*, pages 394–401. Springer-Verlag, 2003.

E. Vidal, F. Thollard, C. de la Higuera, F. Casacuberta, and R. C. Carrasco. Probabilistic finite-state machines-part i. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 27(7):1013–1025, 2005.

A. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13(2):260 –269, 1967.

F. Wessel, R. Schlüter, K. Macherey, and H. Ney. Confidence measures for large vocabulary continuous speech recognition. *IEEE Transactions on Speech and Audio Processing*, 9:288–298, 2001.

F. Wessel and H. Ney. Unsupervised training of acoustic models for large vocabulary continuous speech recognition. *IEEE Transactions on Speech and Audio Processing*, 13(1):23 – 31, 2005.

S. Young et al. *The HTK Book*. Cambridge University Engineering Department, 1995.

X. Zhu. Semi-supervised learning literature survey, 2006.

G. K. Zipf. *Human Behavior and the Principle of Least Effort*. Addison-Wesley (Reading MA), 1949.

CHAPTER 3

Interactive Pattern Recognition

Contents

3.1 Introduction

As said in the Introduction, transcription of old text documents is an expensive and time-consuming task for transcribers. Unfortunately, a fully automatic approach to the transcription problem is currently unfeasible for most applications, as state-of-the-art automatic recognisers cannot still produce acceptable results. An efficient solution is to overcome the problems of both (automatic and manual) approaches by combining them into an interactive approach. The objective of this approach is to employ the benefits of the approaches, i.e. the quality of manual transcription and the efficiency and scalability of an automatic recogniser. This approach has not only been applied to HTR but also to many different areas, in which the system output may not be reliable because of the difficulty or the target usage of the task . For instance, in medical environments, system produced decisions has to be revised by a human operator. As a result, the previously presented probabilistic solution to PR problems, as in our case HTR, has to be extended to include the interaction with humans.

In HTR, the first interactive applications consisted in the simple manual post-processing of an automatically obtained transcription. In these applications, the system propose a transcription to the user, who revise or supervise the system output and corrects it if any errors were found. Similarly, this interactive approach has also been applied on top of other PR systems in very different fields, such as ASR (Luz et al., 2008) or machine translation (MT) (Barrachina et al., 2009), showing improvements over manual approaches. Figure 3.1 depicts a diagram showing the post-editing process for transcribing text documents. First, lines within handwritten text images are automatically recognised by a PR model, next, a user postedits these automatic transcriptions. Refined versions of this process make use of dictionaries, which propose list of similar words to the one that it is being edited, or guide the user on supervising only incorrectly recognised words, in order to improve their effectiveness. However, this approach is only effective under constraint conditions. For instance, the quality of automatic transcriptions has to be acceptable in order for the post-processing to take less time than the manual transcription.

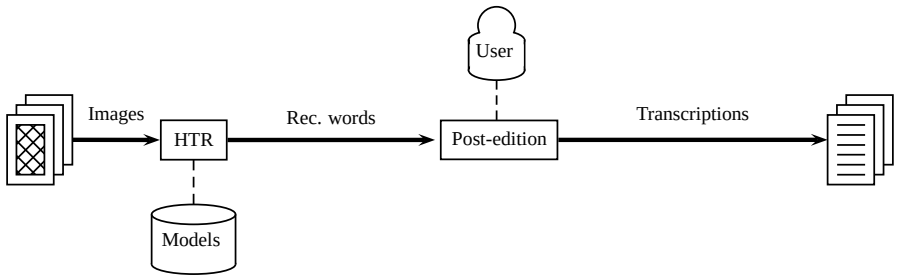


Figure 3.1: Standard post-edition process.

The problem is that, these approaches do not fully exploit the benefits of the interactive approach. For instance, the user interaction applied in correcting the system output gives substantial information that can be used to improve the transcription beyond its simple correction. In fact, they correspond to correct parts on the image, which can be employed on reducing the search space of the recogniser. Hence, it should make the search easier as it

reduces the number of possible transcription to be recognised. In addition, user corrections give information on mis-recognised words by the system. These corrections could be used to improve the current system models improving its future performance. The problem is that the implementation of this two ideas is not easy as existing recognition and training methods has to be modified.

Another important issue is that typically user interactions are applied to correct the system output. However, the whole interactive transcription task includes many different parts besides the automatic recognition and its interactive post-editions. For instance, preprocessing of raw images into suitable feature vectors could be improved employing user interaction to guide certain preprocessing methods. A good example of this process is shown in (España-Boquera et al., 2011), in which users annotate how the ascendants and descendants size of text line images should be normalised. Then, this input is used to train a neural network that performs this process automatically. This process results in an important improvement in terms of recognition accuracy, when compared with unsupervised heuristic methods that perform the same operation. These results rise the fact that it could be better to employ user effort in this way than directly correcting the system output.

In Section 3.2, we present the statistical foundations for interactive transcription. Next, in Section 3.3, it is adapted to the case of interactive transcription of handwritten text document. Last, we also adapt it for the case of interactive document layout analysis in Section 3.4, proving that these approach can be useful whenever an interactive solution is proposed for a task.

3.2 Interactive Pattern Recognition

As said in the Section 2, the pattern recognition problem can be viewed as a probabilistic problem, in which given a feature vector representation of a sample x, its class label y can be obtained by maximising the posterior probability (similarly to Eq. 2.1).

$$\hat{y} = \arg\max_{y} p(y \mid x) \tag{3.1}$$

However, in an interactive approach, a user feedback or interaction f is also available and has to be considered in the classification problem. Adding this variable to the previous equation results in:

$$\hat{y} = \arg\max_{y} p(y \mid x, f) \tag{3.2}$$

which can be decomposed applying the Bayes rule in

$$\hat{y} = \arg\max_{y} p(x \mid y, f)p(y \mid f)p(f) \tag{3.3}$$

The specific modelling of these terms depends on the domain of the user interaction employed.

A simple yet effective consideration is to assume that f has the same domain as y. For instance, in HTR, user interaction may consist in the post-edition of some words y' producing a partial annotation of y. Consequently, the previous equation can be expressed as

$$\hat{y} = \arg\max_{y} p(x \mid y, y')p(y, \mid y')p(y') \tag{3.4}$$

which allows the possibility of employing the same system developed for Eq 3.1, as the feedback is expressed in the same way as the model class y. On the other hand, this simplification can be interpreted as a limitation over the search space (Toselli et al., 2011) because the corrected class helps to reduce the uncertainty of the system in its hypothesis. For instance, a user may specify that some classes are not valid for a particular x, which increases the probability of the others. However, this simplification cannot be always performed, as user interaction may differ in domain or modality with the other variables. For instance, in HTR, user interaction may consist in solving a preprocessing step, which will obtain a feature vector that differs from that of x.

Another important consideration in an interactive task is the possibility of employing the previous system outputs and user interactions. For instance, when transcribing a handwritten text document, the transcriber will sequentially annotate each line. This information can be used to improve the system and so, the quality of the transcriptions produced. Considering the interactive transcription of N handwritten text images x_1^N and their corresponding text transcriptions y_1^N, Eq 3.2 is expressed as:

$$\hat{y}_1^N = \arg\max_{y_1^N} p(y_1^N \mid x_1^N, f_1^N) \tag{3.5}$$

which by applying the chain rule of probability can be viewed as a sequential process

$$\hat{y}_1^N = \arg\max_{y_1^N} p(y_1 \mid x_1, f_1) p(y_2 \mid x_1^2, f_1^2, y_1) \cdots p(y_N \mid x_1^N, f_1^N, y_1^{N-1}) \tag{3.6}$$

As observed, posterior probabilities now depend on all previously seen variables. So all the possible classification combinations should have to be considered, which may be unfeasible for many applications. This problem can be tackled by making some assumptions. It can be assumed that the classification of the actual sample is independent from the previous one, which will result in a sequential recognition process while adapting from all previously classified samples. Alternatively, it can be assumed a n-order dependency on only the previous n-ones, thus, reducing the resulting model complexity.

3.3 Interactive Handwriting Recognition

As said in Section 2.6, the objective of interactive handwriting recognition is the efficient usage of the user effort available obtaining the best transcription possible of document. This objective would be achieved by correctly modelling and applying Eq. 3.6, but its direct estimation and search is unfeasible. However, it gives an idea of how the best search could be achieved. Eq. 3.3 can be transformed to fit the case of interactive HTR. Given a feature vector $\mathbf{x}$ and a user interaction f, the most probable transcription $\mathbf{w}$ can be obtained by:

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} p(\mathbf{x} \mid f, \mathbf{w}) p(\mathbf{w} \mid f) p(f) \tag{3.7}$$

where considering that all user interactions are equally likely $\forall f : p(f) = \frac{1}{|f|}$, and there the previous equation becomes

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} p(\mathbf{x} \mid f, \mathbf{w}) p(\mathbf{w} \mid f) \tag{3.8}$$

Up to this point there has not been any decision on how the user interaction is performed. In this thesis, user interaction is considered the supervision and correction (if necessary) of some recognised words. As the effort is limited, only a handful of words would be supervised. This creates an additional problem, how to select which words are going to be supervised. In fact, the supervision of some words may help the system more than others. For instance, the supervision of a correct word would waste the user effort as they were correctly recognised by the system, while the supervision of an incorrect word improves the transcription and adds new annotated data. This problem is studied by a subfield of PR called Active Learning (AL) (Settles, 2010). AL studies how to select which unannotated samples are to be supervised so that their supervision maximises the performance of the system. Including AL in an interactive HTR approach leads to the so called guided post-editing transcription, as depicted in Figure 3.2. As observed, using an automatically generated transcription, AL techniques select words with low confidence, that is words that are likely to be incorrect, and ask the user to supervise them.

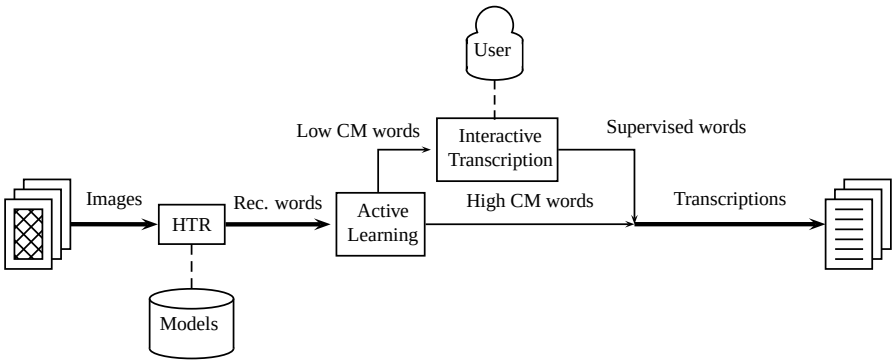


Figure 3.2: Guided post-edition interactive process

In Figure 3.2, we can also observe that, differently from the previous diagram in Figure 3.1, the system would only ask the user to supervise some words. A good example of this approach for HTR was studied in (Tarazón et al., 2009). In this work, an accurate error detection method could detect most of the errors by only supervising few recognised words, skipping most correct words.

Once a few words are supervised by the user, the resulting transcription is improved. However, this supervision f can be used within Eq. 3.8 to further improve the transcription beyond its simple correction. This operation will be referred as constrained search, as it corresponds to the search of the optimum transcription constrained to the user interaction (or requirements). This new search will obtain better results as user supervision reduces the uncertainty of the system, guiding the search towards better hypotheses.

Figure 3.3 shows the addition of this step to the latter diagram. As observed, this new step is performed after user supervision. The estimation and implementation of this step strongly depends on the user interaction considered. For instance, Toselli et al. (2007) employed constrained search for HTR in which user interaction represents the correction of prefixes. This is motivated by the interactive approach followed, in which users continuously supervise

the system output up the first incorrect word. Then, the system recomputes its hypothesis, and the user continues the supervision. Similarly, Kristjannson et al. (2004) applied the described constrained search to the task of interactive information extraction. In the case of this thesis, user interactions can correspond to the correction of random words within the image, independently from its position. This method is fully detailed later in Section 5.4.1.

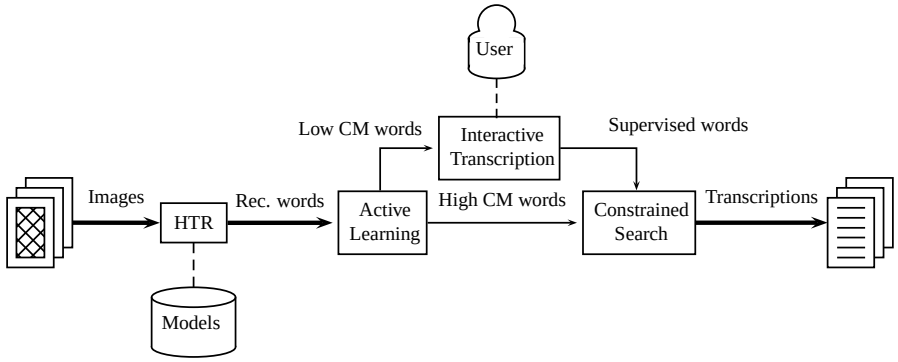


Figure 3.3: Constrained search after a guided post-edition process

Hitherto we have described how to obtain the transcription of a given image and some user interactions. However, only the transcription of single line image has been considered until now, while typically a whole set of lines would be transcribed. Extending the same assumptions we have made for a single image to the case of N lines, we obtain

$$\hat{\mathbf{w}}_1^N = \arg\max_{\mathbf{w}_1^N} \prod_{n=1}^{N} p(\mathbf{x}_n \mid \hat{f}_1^n, \mathbf{w}_1^n, \mathbf{x}_1^{n-1}) p(\mathbf{w}_1^n \mid \hat{f}_1^n, \mathbf{x}_1^{n-1}) \tag{3.9}$$

which is similar to the previous Eq. 3.8 but depending on all previous images, specifically

$$\begin{aligned}\hat{\mathbf{w}}_1^N = \arg\max_{\mathbf{w}_1^N} \prod_{n=1}^{N} & p(\mathbf{x}_n \mid \hat{f}_n, \mathbf{w}_n, \hat{f}_1^{n-1}, \mathbf{w}_1^{n-1}, \mathbf{x}_1^{n-1}) \\ & p(\mathbf{w}_n \mid \hat{f}_n, \hat{f}_1^{n-1}, \mathbf{w}_1^{n-1}, \mathbf{x}_1^{n-1}) p(\mathbf{w}_1^{n-1} \mid \mathbf{x}_1^{n-1}, \hat{f}_1^n)\end{aligned} \tag{3.10}$$

However, the search proposed in the latter term presents two main problems. First, the search itself that has to be performed globally for all possible transcription of all images, and last, the difficulty of estimating the terms due to the dependence on several variables. The first problem can be avoided by considering that the transcription process is performed sequentially, one line after the other

$$\hat{\mathbf{w}}_1^N = \left\{ \arg\max_{\mathbf{w}_n} p(\mathbf{x}_n \mid \hat{f}_n, \mathbf{w}_n, \hat{f}_1^{n-1}, \mathbf{w}_1^{n-1}, \mathbf{x}_1^{n-1}) p(\mathbf{w}_n \mid \hat{f}_n, \hat{f}_1^{n-1}, \mathbf{w}_1^{n-1}, \mathbf{x}_1^{n-1}) \right\}_{n=1}^{N} \tag{3.11}$$

which in the case of the application of this thesis is natural, as it is assumed that this order is given by the natural sequential structure of handwritten text documents. This assumption

is fair for transcription experts, because they typically transcribe the document from the beginning to the end. Alternative methods of order selection are out of the scope of this thesis. Interested readers are referred to active learning literature (Settles, 2010), where this matter is studied in detail.

Finally, only the problem of multiple dependencies remains. However, as it can be observed in Eq. 3.11, previous images, transcriptions and user interactions have been already produced, and thus these dependencies can be incorporated as new annotated images for the model parameters estimation. As in the training of HMMs presented in Eq. 2.9, the model parameters Θ of an interactive recognition system can be estimated by maximising the likelihood function L over a given set of annotated samples S. Given an interactive recognition task, we assume that model parameters are estimated with all annotated data available. For instance, the model parameters when recognising the sample n are estimated as:

$$\Theta^{(n)} = \arg\max_{\Theta} L(\Theta; S \cup \{\mathbf{x}_1^{i-1}, \mathbf{w}_1^{i-1}, f_1^{i-1}\}) \tag{3.12}$$

then, we approximate then dependencies on Eq. 3.11 as:

$$\hat{\mathbf{w}}_1^N = \left\{ \arg\max_{\mathbf{w}_n} p_{\Theta^{(n)}}(\mathbf{x}_n \mid \hat{f}_n, \mathbf{w}_n) p_{\Theta^{(n)}}(\mathbf{w}_n \mid \hat{f}_n) \right\}_{n=1}^{N} \tag{3.13}$$

In this approximation, each time a transcription is recognised, model parameters Θ are updated using all available data, so recognition of posterior images improves. In fact, this step would be performed as a complete re-training of all models. However, due to the high computational cost of the process, as it implies retraining image and language models, it is typically only performed once a set of new images have been transcribed. This process can also be carried out in an on-line fashion, in which model parameters are updated with each sample (Ortiz-Martínez et al., 2010). In this thesis, we have centered on the first, complete re-training, as it obtains equal or better results than an on-line estimation. Furthermore, we also study the adaptation of partially supervised transcription, as the transcription obtained could have not been completely supervised by an user, and thus, errors could remain.

Concluding, once this model adaptation and retraining have been performed, the interactive transcription process corresponds to the one defined in Figure 3.4. In this figure, we can observe how each line is processed. First, user supervises some words, which are used to further improve the transcription by constrained search, and finally improve the HTR system by re-training, therefore improving the recognition of posterior lines.

3.4 Interactive Document Layout Analysis

In this section, we present another application of interactive pattern recognition for the task of document layout analysis (DLA). This task corresponds to the first step of the whole transcription process of handwritten text pages. In this step, the location of the text to be transcribed is detected, as document pages may include different layouts, such as multi-column texts. A correct annotation of the layout will help the system to correctly process the posterior steps, such as text baseline detection, and thus improve the final automatic recognition.

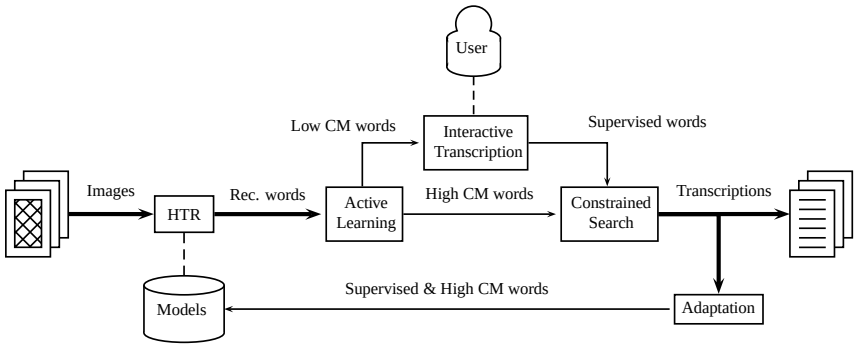


Figure 3.4: Interactive transcription process, in which adaptation is applied using the resulting partially supervised transcriptions

Typically, this problem is addressed as a syntactic analysis problem (Mao et al., 2003), in which the document layout of a page is represented as a logical relation among the components of the page, i.e. text block, images, captions, etc. When applied to the transcription of handwritten old text documents, an interactive approach to this task would be adequate, as at the beginning there is little annotated information to train a reliable system. Intelligent interaction will help the user to annotate the document layout more precisely, while improving the HTR performance.

Given a document image represented by a feature vector z, its document layout structure h has to be detected. Variable h is divided in two variables, l and s, which represent its layout contour and class, respectively. This detection problem can be solved using a statistical PR approach following Eq.3.1

$$\hat{h} = \underset{h=(l,s)}{\arg\max}\, p(z \mid l, s)p(l, s) \tag{3.14}$$

in which the layout structure h is obtained by maximising its posterior probability given z.

Then, applying an interactive approach in which some user interactions f have been performed, the latter variable is introduced in the search as explained in Section 3.2

$$\hat{h} = \underset{h=(l,s)}{\arg\max}\, p(z \mid \mathbf{f}, l, s)p(\mathbf{f} \mid l, s)p(l \mid s)p(s) \tag{3.15}$$

in which the term $p(\mathbf{f} \mid l, s)$ deals with the probability of the applied user interactions given the possible layouts. This probability is the most important part as it guides the conventional search.

The proposed search was applied in (Ramos-Terrades et al., 2010) in the detection of the layout of a handwritten text document. In this task, the layout was formed by square contours l and there were only to two types of classes, “front” and “back”, which corresponds to the left and right pages, respectively. Empirical results were obtained varying the size of the available user interaction history. Results showed that even when dealing with simple layout, a few user interactions can result in a great improvement of the overall process.

3.5 Conclusions

In this chapter, a statistical approach to interactive pattern recognition has been presented. This approximation introduces two main features, the dependency of all model terms with user interactions, and the dependence on previous system recognitions. Due to the unfeasibility of a direct approach, several simplifications have to be performed while trying to maintain the original model complexity. In addition, as this simplifications depend on the task in which these approach is applied, two different applications were presented for the tasks of interactive handwritten text recognition and interactive document layout analysis.

This latter application has led to a publication in an international conference:

- O. Ramos, **N. Serrano** and A. Juan. Interactive-predictive detection of handwritten text blocks. In *Proceedings of the 17th Document Recognition and Retrieval Conference (DRR 2010)*. San Jose (USA). January 2010.

Bibliography

S. Barrachina, O. Bender, F. Casacuberta, J. Civera, E. Cubel, S. Khadivi, A. L. Lagarda, H. Ney, J. Tomás, and E. Vidal. Statistical approaches to computer-assisted translation. *Computational Linguistics*, 35 (1):3–28, 2009.

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

T. Kristjannson, A. Culotta, P. Viola, and A. McCallum. Interactive information extraction with constrained conditional random fields. In *Proc. of the 19th National Conference on Artificial Intelligence (AAAI 2004)*, pages 412–418, 2004.

S. Luz, M. Masoodian, and B. Rogers. Interactive visualisation techniques for dynamic speech transcription, correction and training. In *Proc. of the 9th Int. Conf. on Human-Computer Interaction (CHINZ 2008)*, pages 9–16, Wellington, New Zealand, 2008.

S. Mao, A. Rosenfeld, and T. Kanungo. Document structure analysis algorithms: a literature survey. In *Proc. of SPIE-IS&T Electronic Imaging (DDR X)*, pages 197–207, 2003.

D. Ortiz-Martínez, I. García-Varea, and F. Casacuberta. Online learning for interactive statistical machine translation. In *Proceedings of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL HLT)*, pages 546–554, 2010.

O. Ramos-Terrades, N. Serrano, A. Gordó, E. Valveny, and A. Juan. Interactive-predictive detection of handwritten text blocks. In *Proc. of SPIE-IS&T Electronic Imaging (DDR XVII)*, pages 75340Q–(1–10), 2010.

B. Settles. Active learning literature survey. Technical report, 2010.

L. Tarazón, D. Pérez, N. Serrano, V. Alabau, O. Ramos-Terrades, A. Sanchis, and A. Juan. Confidence measures for error correction in interactive transcription of handwritten text. In *Proc. of the 15th Int. Conf. on Image Analysis and Processing (ICIAP 2009)*, pages 567–574, Vietri sul Mare (Italy), 2009.

A. Toselli, V. Romero, L. Rodríguez, and E. Vidal. Computer assisted transcription of handwritten text. In *Proc. of the 9th Int. Conf. on Document Analysis and Recognition (ICDAR 2007)*, pages 944–948, Curitiba (Brazil), 2007.

A. H. Toselli, E. Vidal, and F. Casacuberta, editors. *Multimodal Interactive Pattern Recognition and Applications*. Springer, 2011.

CHAPTER 4

Annotation of Handwritten Text Documents

Contents

4.1 Introduction

In this chapter, we describe the digitisation and annotation process of two real handwritten text documents called GERMANA and RODRIGO. The documents were carefully selected to serve as benchmarks of interactive transcription approaches, that is the main topic of this thesis. The task of automatically transcribing these two document is not straightforward, as the scarcity of external resources make complex building a good PR system. The annotation of this two documents has produced two databases of similar size to standard database for HTR, such as the IAM database (Marti and Bunke, 2002). Moreover, the databases have been made freely available for research purposes to facilitate empirical comparison of different approaches to document layout analysis, text line extraction and off-line handwriting recognition.

In addition, we also present the sequential process to create a baseline system from fully supervised transcription. In this process, we tuned the feature extraction method and the HTR system parameters, as well as we applied some tools to reduce the language complexity. Results are discussed on each step on the validation set of each corpus using only a small part of the document as training. Next, the best system obtained is used to sequentially transcribe the remainder chapters, as it would be performed in a post-edition approach. Specifically, each chapter is automatically transcribed, then fully revised by a user, and finally added to the training set. This experiment will serve to observe the overall difficulty of these tasks.

This chapter is divided in three sections. Firstly, GERMANA and RODRIGO are described in Section 4. Secondly, baseline experiments are thoroughly described in Section 4.3. Finally, conclusions and future work are reviewed in Section 4.4.

4.2 Annotation of GERMANA and RODRIGO

4.2.1 GERMANA

GERMANA is the result of digitising and annotating a 764-page Spanish manuscript entitled *"Noticias y documentos relativos a Doña Germana de Foix, última Reina de Aragón"*[a] and written in 1891 by Vicent Salvador, the Cruïlles' marquis. The original manuscript is preserved in the Nicolau Primitiu Collection at the Valencian Library (BiValDi). Manual transcription of GERMANA is not a particularly difficult task for several reasons. First, it is a single-author book on a limited-domain topic: the life of *Germana de Foix*[b] (1488-1538), niece of King Louis XII of France and second wife of Ferdinand the Catholic of Aragon. Also, the original manuscript was well-preserved and most pages only contain nearly calligraphed text written on ruled sheets of well-separated lines.

It goes without saying that text line extraction and off-line handwriting recognition on GERMANA is not, by contrast, particularly easy. GERMANA has typical characteristics of historical documents that make things difficult: spots, writing from the verso (or front pages) appearing on the recto (or back pages), unusual characters and words, etc. Also, the manuscript includes many notes and appended documents that are written in languages dif-

[a] In English, "Related news and documents of Mrs Germana of Foix, last queen of Aragón

[b] Her biography can be found at `http://wikipedia.org/wiki/Germana_de_Foix`

ferent from Spanish, namely Catalan, French and Latin. All in all, we think that GERMANA entails an appropriate trade-off between task complexity and amount of data. To our best knowledge, it was at its publication date, the first publicly available database for handwriting research, mostly written in Spanish and comparable in size to standard databases such as IAM. Due to its sequential book structure, it is also well-suited for realistic assessment of *interactive* handwriting recognition systems, in which a user follows a sequential process to transcribe the document from the beginning to the end. Moreover, it can be used as well to test approaches for language identification and adaption from single-author handwriting.

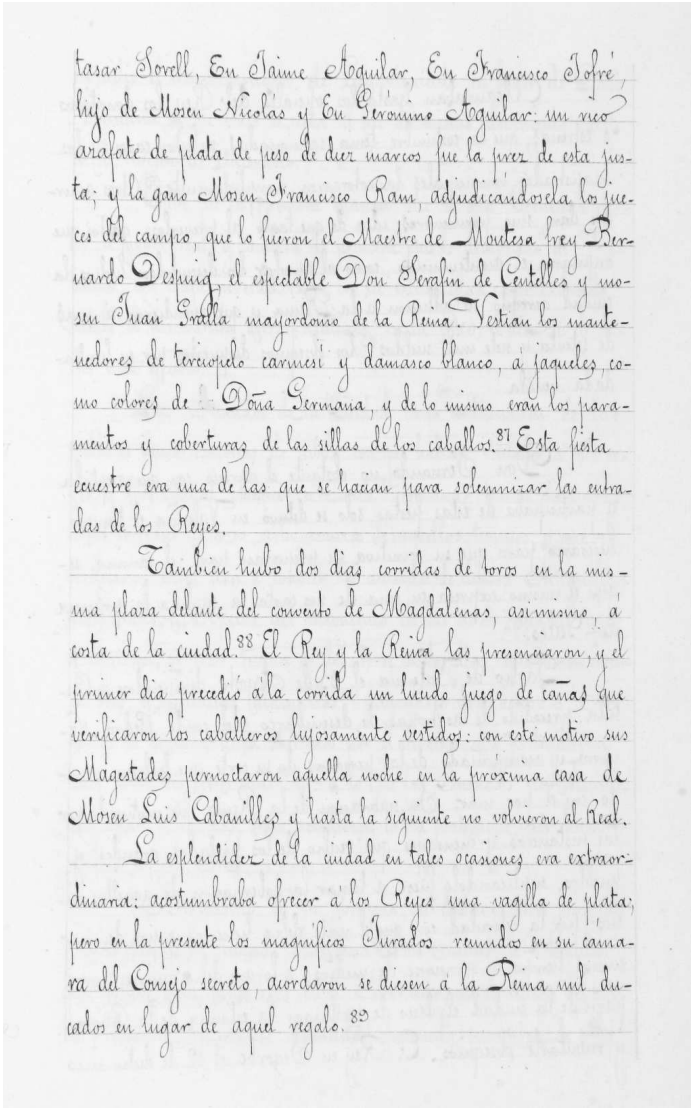
GERMANA manuscript is divided into into 17 sections. However, for simplicity, we will distinguish only 7 parts of the manuscript:

1. *Front matter (pp* 1–6*):* a half title, a title and a portrait of *Doña Germana de Foix*.
2. *The chapters (pp* 7–180*):* 174 pages divided into 6 chapters, each one devoted to a distinct period in the life of Germana.
3. *Notes (pp* 181–282*):* 290 numbered notes referenced in the chapters.
4. *Biography notes (pp* 283–302*)* of 8 relevant persons mentioned in the second part.
5. *Documents (pp* 303–540*):* handwritten copies of 71 historical documents related to the life of Germana.
6. *Illustrations (pp* 541–716*):* 4 documents with their own notes appended at the end.
7. *Back matter (pp* 717–764*):* various indices and images.

Most pages only contain handwritten text aligned to horizontal rules in a simple template of either 24 (pp 1–180 and 729–764) or 32 (pp 181–728) lines. As an example, the page 67 is shown in Figure 4.1. Note that the handwriting is easily readable and tightly aligned to horizontal rules.

The manuscript is solely written in Spanish up to page 180. After this page, however, the reader can also find text in Catalan, French, Latin and, to a lesser extent, German and Italian. In the third part, there are 33 notes (mostly) written in Catalan (4, 47, 50, 73, 78, 79, 81, 82, 84, 85, 87-91, 94-96, 134, 177, 194, 205, 209, 214, 227, 229, 236, 238, 261, 266-268 and 270); 18 in French (1, 2, 15, 22, 23, 25, 29, 44-46, 71, 109, 110, 119, 155, 170, 257 and 280); and 1 in German (180). Also, there are 24 documents in the fifth part that are written in Catalan (7, 8, 27, 29, 31-33, 36-40, 44, 48-54, 59, 64, 68 and 69); 10 in Latin (2, 4-6, 12, 24, 34, 42, 43, 70); 1 in French (7); 1 in German (25); and 1 in Italian (65). Biography notes and Illustrations are primarily written in Spanish, though there is also some content in Catalan (a short excerpt of 13 lines starting at the last line on page 300; notes 39, 47 and 61 of illustration C; and note 17 of illustration D).

The manuscript was carefully scanned by experts from the Valencian Library at 300dpi in true colours. As with historical documents in general, scanned pages have noise effects like spots, tears, ink fading and transparency of back side. Also, they show a slight warping due to book binding. Nevertheless, the manuscript can be easily read and thus we decided not to apply any preprocessing to it for the purpose of annotating ground-truth. Ground-truth annotation of layout of GERMANA consisted of two parts. On the one hand, all text blocks



tasar Sorell, En Jaime Aguilar, En Francisco Jofré, hijo de Mosen Nicolás y En Gerónimo Aguilar: un rico arafate de plata de peso de diez marcos fue la prez de esta justa; y la ganó Mosen Francisco Ram, adjudicándosela los jueces del campo, que lo fueron el Maestre de Montesa frey Bernardo Despuig, el espectable Don Serafin de Centelles y mosen Juan Ensalla mayordomo de la Reina. Vestian los mantenedores de terciopelo carmesí y damasco blanco, á jaqueles, como colores de Doña Germana, y de lo mismo eran los paramentos y coberturas de las sillas de los caballos.[87] Esta fiesta ecuestre era una de las que se hacian para solemnizar las entradas de los Reyes.

También hubo dos dias corridas de toros en la misma plaza delante del convento de Magdalenas, asimismo, á costa de la ciudad.[88] El Rey y la Reina las presenciaron, y el primer dia precedió á la corrida un lucido juego de cañas que verificaron los caballeros lujosamente vestidos: con este motivo sus Magestades pernoctaron aquella noche en la próxima casa de Mosen Luis Cabanilles y hasta la siguiente no volvieron al Real.

La esplendidez de la ciudad en tales ocasiones era extraordinaria: acostumbraba ofrecer á los Reyes una vagilla de plata; pero en la presente los magníficos Jurados reunidos en su cámara del Consejo secreto, acordaron se diesen á la Reina mil ducados en lugar de aquel regalo.[89]

Figure 4.1: Page 67 of GERMANA.

were marked with minimal enclosing rectangles and, within each text block, each text line was marked by its (straight) baseline. This was done semi-automatically by means of the GIDOC prototype (see Appendix A). All blocks and baselines detected automatically were also manually supervised, and corrected when needed.

On the other hand, the whole manuscript was transcribed line by line, by paleography experts. The transcription process did not start from scratch, but from a partial transcription

produced by experts from the Valencian Library during 2002. This partial transcription covered most of the manuscript (76%), but it was not directly applicable to handwriting research, mainly because it did not include original page and line breaks. Therefore, to produce the final transcription, this partial version was first reviewed and then completed. It was done again by paleography experts, in accordance with the following transcription rules:

- Page and line breaks are copied exactly.
- Blank space is only used to separate words.
- No spelling mistakes are corrected.
- No case or accentuation change is done.
- Punctuation signs are copied as they appear.
- Word abbreviations are first copied verbatim, except for subindices and superindices, which are written in LaTeX-like notation as `_{sub}` and `^{super}`, respectively. Then, they are followed by the corresponding word between brackets. Thus, for instance, `D`$^{\text{a}}$`.` is transcribed as `D^{a}.[Doña]`. Figure 4.2 show an examples of an abbreviation and a superindex. This special annotation will be used to build the language model part of the HTR system.

Figure 4.2: Example of a line with abbreviations and superindexes.

Also, to facilitate language-dependent processing of the manuscript, each transcribed line was manually labelled in accordance with its dominant language. The total time required for a single expert to manually transcribe the whole manuscript was estimated as 232 hours; that is, approximately 30 minutes per page on average. Note that, the time require to mark the text block and its baselines is also included.

Table 4.1 contains some basic statistics drawn from the transcriptions. It must be noted that, these statistics include some pages that cannot by used for HTR, as they contain graphics or genealogical trees. The amount of data used for each experiment will be described in the experiments Section 4.3 . These statistics were computed after applying the following preprocessing steps:

1. Substitution of abbreviations by their corresponding words.
2. Concatenation of hyphenated words at line ends with their remainders.
3. Isolation of punctuation signs.

Table 4.1: Basic statistics of GERMANA

GERMANA						
				Lexicon		
Language	Pages	Lines	Words(K)	Size(K)	Singletons(%)	Char. set size
Spanish	595	16599	176.8	19.9	55.6	111
Catalan	87	2417	26.9	4.6	63.2	86
Latin	29	951	8.3	3.4	69.2	87
French	8	266	3.0	1.1	71.1	82
German	8	228	1.5	0.6	52.7	71
Italian	2	68	0.8	0.3	67.3	59
None	35	0	0.0	0.0	0.0	0
All	711	20150	225.3	28.8	58.7	115

Note that the Spanish part of GERMANA comprises about $17K$ text lines and $177K$ running words from a lexicon of $20K$ words, which is comparable in size to standard databases. It is also worth noting that 56% of the words only occur once (singletons). The database is available at the PRHLT website[c] for non-commercial research. It contains approximately $21K$ text lines that comprises about $217K$ running words from a vocabulary of $30K$ words which, apparently, is a reasonable amount of data for single-author handwriting and language modelling. The interested reader is referred to (Belenguer, 2007) for a deep study of the manuscript from a historian's point of view, and for a printed transcription of the manuscript though, as it was not intended for handwriting research, it was reformatted for better readability.

4.2.2 RODRIGO

RODRIGO is a manuscript from 1545 entitled *"Historia de España del arçobispo Don Rodrigo"*, and completely written in old Castilian (Spanish) by a single author. It is a 853-page bound volume divided into 307 chapters describing chronicles from the Spanish history. Most pages only contain a single text block of nearly calligraphed handwriting on well-separated lines. Its size is similar to the previously described database, GERMANA, and they also present some common features, such as homogeneous writing or the presence of an unique block per page. Its first part was copied from an older (XV century) manuscript, followed by an addition of posterior chronicles. The original manuscript is preserved in the "Castilla de la Mancha" library (Bib). As in GERMANA, handwritten lines are easily readable and tightly aligned, containing 24 lines on average. According to experts, the manuscript writing style corresponds to Humanistic script, similar to the Italic script (Millares and Ruiz, 1983) but with textual Gothic influences. As an example, pages 15 and 16 are shown in Figure 4.3.

Other characteristic details of RODRIGO that can be clearly appreciated in Figure 4.3 are:

- The author tends to embellish the writing, specially in broad white spaces, resulting in

[c] `http://prhlt.iti.es`

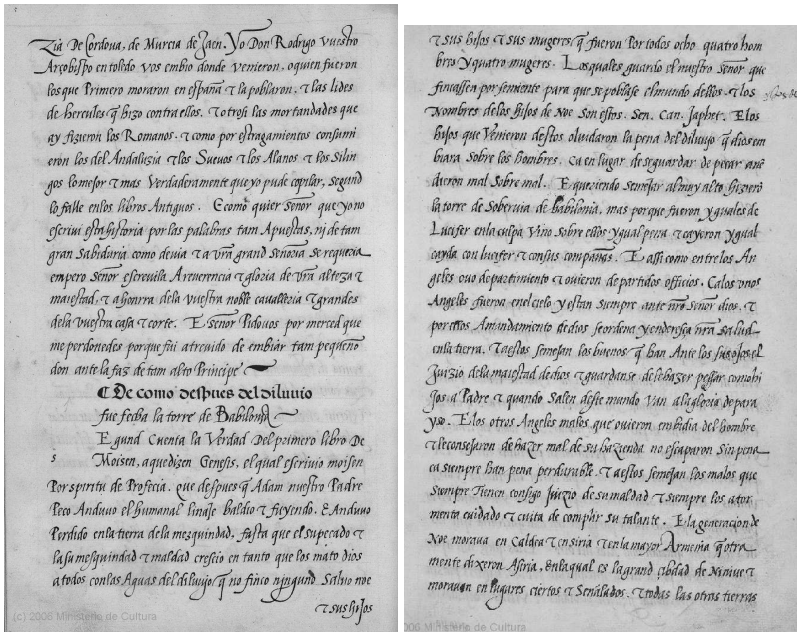


Figure 4.3: Pages 15 and 16 of RODRIGO.

the extension of some ascenders and descenders across whole words.

- Natural blank spaces between successive words are often omitted; e.g., the words "de la" are written as a single word "dela" in the third line from the bottom of page 15. Sometimes, on the contrary, artificial blank spaces are inserted within a single word; e.g., the word "llegaronse" is written as two words, "llegaron se".

- Each chapter should begin with a dropcap, but the manuscript contains no dropcaps, probably because it was never brought to an artist to do so. Instead, there is a blank area in each position where a dropcap should have been inserted and, in most cases, the corresponding letter is written in small size.

- The first words in each even page are also copied in the bottom right corner of its preceding page.

- There is no indicator at the end of the line when a word is splitted.

On one hand, the manuscript was carefully digitised by experts from the Spanish *Ministry of Culture*, at 300dpi in true colours, and it is publicly available at RodrigoMCU. As with historical documents in general, scanned pages have noise effects like spots, tears, ink fading and transparency of back side. Also, they show a slight warping due to book binding. Nevertheless, the manuscript can be easily read and thus we decided not to apply any preprocessing to it (apart from de-saturation) for ground-truth annotation. Next, we followed

an annotation procedure very similar to the one used for the GERMANA database. First, all text blocks were annotated with minimal enclosing rectangles and, within each text block, each text line was marked by its (straight) baseline by means of the GIDOC prototype. All detected blocks and baselines were also manually supervised, and corrected when needed.

On the other hand, the whole manuscript was transcribed line by line, by a paleography expert, in accordance to transcription rules in GERMANA and three more new rules:

- Missing natural blank spaces between successive words are indicated by the symbol "◡".
- Inserted artificial blank spaces within words are indicated by the symbol "␣".
- The symbol "$" is appended to each line having a broken word at its end.

The total time required for a single expert to manually annotate (text blocks, baselines and transcriptions) the whole manuscript was estimated as 500 hours; that is, approximately 35 minutes per page on average. The complete annotation of RODRIGO is publicly available, for non-commercial use, at the PRHLT website[d] . It comprises about $20K$ text lines and $231K$ running words from a lexicon of $17K$ words. It is worth noting that more than half of the words in the lexicon (54.4%) are singletons (or *hapax legomena*), but they only account for a 4.1% of the running words. Please see Table 4.2 for some basic statistics. It must be noted that, statistics were drawn from the transcriptions in accordance to the rules applied in the computation of statistics in GERMANA.

Table 4.2: Basic statistics of the RODRIGO text transcriptions (with isolated punctuation signs and abbreviations substituted by their corresponding words). Perplexity was computed using a bigram language model and a 100-fold cross-validation experiment. Singletons refers to words occurring exactly once.

RODRIGO	
Pages	853
Lines	20357
Running words	$232K$
Perplexity	166
Lexicon size	$17.3K$
Singletons (%)	54.4
Character set size	115

4.3 Baseline Experiments

In this section, we describe all experiments that have been performed to obtain the baseline system used in the interactive approach. As we are dealing with the interactive transcription, first, an HTR system has to be built from scratch for the subsequent interactive experiments. We divided both documents, GERMANA and RODRIGO, in blocks of one thousands

[d] `http://prhlt.iti.es`

lines, except for the first that was splitted into two blocks of five hundred lines, and the last blocks, which also contains the last remnant of lines. We consider that the first two blocks are manually annotated to build the initial baseline system. Concretely, the first block is used as training and the second block as validation. Table 4.3 shows basic statistics of training and validation blocks. It must be noted that, the number of words were calculated directly from the paleography reference, in which no punctuation marks is isolated from the closed words. Differently, the size of the lexicon is calculated from the transcription once it is appropriately parsed. The size of lexicon in the validation set is expressed as the number of new words added to the vocabulary, which corresponds to the number of Out-Of-Vocabulary words (OOVs). As observed, the size of the proposed partitions is small and it is not sufficient to estimate all the parameters of an HTR system. However, data scarcity is one of the problems that is tackled in this type of tasks.

Table 4.3: Statistics of the first and second blocks in GERMANA and RODRIGO using the reference transcriptions.

	GERMANA		RODRIGO	
	Train	Validation	Train	Validation
Lines	500	500	500	500
Words	4658	5034	5538	5507
Lexicon	1973	+1567	1812	+1033
OOVs(%)	-	36.5	-	23.6

As described, there is a high number of parameters and methods involved in the development of HTR systems. For sake of simplicity and due to the high cost (human and computational) required to adjust all parameters, we have considered some steps to be common for all experiments. For instance, both databases share the same preprocessing. Line images were preprocessed as follows. First, a pixel value normalisation was applied to denoise the images. Second, the writing slant was corrected. Finally, script ascendants and descendants were normalised. Similarly, all HTR systems were trained using the same toolkits. The image model was modelled as a HMM and it was trained using the AK toolkit (Giménez, 2011). Pararelly, the LM model was modelled as interpolated bigrams using the modified Knesser-Ney discount (Chen and Goodman, 1999), and it was estimated by means of the SRILM toolkit (Stolcke, 2002). The selection of this concrete LM is caused by the fact that there is not sufficient data to compare different LMs, and this model has been shown to perform well in other related works (Bertolami and Bunke, 2008). We consider this closed set of parameters and methods to be fixed, as their fine tuning is expected to not improve the system as much as the other parameters.

In the remainder, the quality of each experiment is measured in terms of Word Error Rate (WER). WER is the average number of editing operations to convert the recognised transcription into the reference transcription, divided by the number of reference words. It must be noted that, as editing operations include the insertion operation, WER is not a percentage and it could surpass the 100%, as the system may recognise more words than there are in the reference.

4.3.1 Basic Parameter estimation

The first step to build the baseline system is the correct estimation of HMM parameters. As said in the Chapter 2, each character is modelled as a HMM of n number of states, in which each state models a mixture of gaussians of g components. In the first experiment, we evaluated the recogniser performance in the validation block for different values of n and g when training the HMMs from the training block. The tuning of the number of states typically starts as the mean number of feature vectors per character. Alternatively, the number of components per mixture starts in one, and it is continuously doubled until the recognition results stop improving. Concretely, we performed the described experiments in both database for $n = \{2, 3, 4, 5\}$ and $g = \{1, 2, 4, 8, 16, 32, 64, 128\}$. In addition to these parameters, we have to tune another two important parameters that balance the score from the HMM with the scores from the LM. These parameters are included as weights in the estimation of the best transcription given in Eq. 2.3

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} p(\mathbf{x} \mid \mathbf{w}) p(\mathbf{w})^{\alpha} \beta \tag{4.1}$$

where α is commonly known as Grammar Scale Factor (GSF), whereas β is typically referred to as Word Insertion Penalty (WIP). These two parameters are used to correctly scale both models, HMMs and LM, as their scores are typically in different magnitude ranges. Many different values for GSF and WIP were tested because, on the contrary to n and g, their modification do not imply to fully re-train the system.

The feature extraction method used in these experiments is the derivative-based described in (Pastor, 2007). The n-gram language model is trained from paleographic transcriptions directly, without any kind of parsing. All words from the training block are added to the lexicon, in which each word is transliterated as its corresponding characters. It must be noted that as opposite to ASR, in HTR the transliteration is not ambiguous. However, we consider two different entries for each word, with and without an ending blank character at the end. The motivation of this double representation is to deal with words at the end of the line and also word overlapping, in which there is not a real blank in the image.

Figure 4.4 shows the results in terms of WER of the parameter optimisation in the validation block. As observed, results vary significantly depending on the parameters. In GERMANA, one of the best recognition results is 61.14%, which is achieved with 4 states per character of 32 mixture components each. Even though that there are better results using a higher number of states, there are not statistically significant differences between them, according to a bootstrap evaluation (Efron and Tibshirani, 1994). Consequently, we have selected the described system because is obtained with a smaller number of states. Similarly, in RODRIGO, the best significant result, 48.76% of WER, that employs the smallest number of states is obtained with 4 states of 32 mixture components each. From these initial results, it can be observed that the transcription of GERMANA is a harder task than RODRIGO.

4.3.2 Punctuation marks isolation

Transcriptions showed that OOVs account for most of the errors. As observed in Table 4.3, OOVs ratio is quite high in the validation block, specially in GERMANA. A solution to this problem is to take advantage from the orthographic rules of the punctuation signs, i.e.

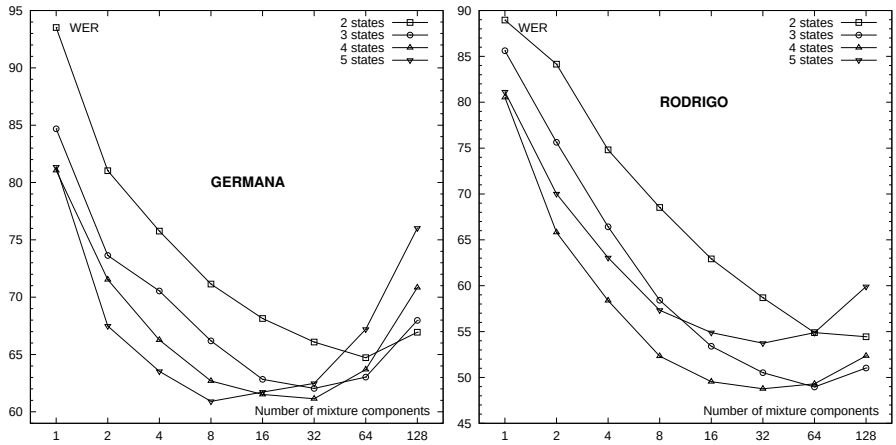


Figure 4.4: Recognition results on GERMANA and RODRIGO when varying the number of states and mixture components.

Table 4.4: Statistics of the first and second blocks in GERMANA and RODRIGO when isolating the punctuation signs.

	GERMANA		RODRIGO	
	Train	Validation	Train	Validation
Lines	500	500	500	500
Words	5205	5617	6006	5783
Lexicon	1816	+1336	1625	+892
OOVs(%)	-	29.3	-	20.2

punctuation marks are concatenated to the previous word and they are followed by a blank character. Punctuation mark may create OOVs and difficult the LM estimation. For instance, the word “arbol” in the validation block is not recognised because it is not present in the lexicon. However, the word “arbol,”, which is identical but for the comma, is in the lexicon. We can solve this problem by defining a set of punctuation marks to be isolated, reducing the number of OOVs while improving the LM estimation. After a line is recognised, all the recognised punctuation marks are concatenated with the previous word, in order to compare the transcription with the paleographic reference.

Table 4.4 shows the statistics of train and validation blocks after the isolation of punctuation marks has been applied. As observed, the number of words increases while the size of lexicon decreases, which results in a reduction of the lexicon and thus, the search space. In addition, it also improves the estimation of the LMs, as there are more words to train a smaller lexicon.

We also repeated the experiment to adjust the number of states and mixture components but isolating the punctuation marks on the train block, because the best parameters from the previous experiment, may not be the best parameters under this new approach. The results

Table 4.5: Results in GERMANA and RODRIGO when considering the isolation of special symbols.

	Baseline	Symbol isolation
GERMANA	61.14	54.07
RODRIGO	48.76	45.41

comparing both approaches, *Baseline*, which corresponds to the best previous system; and *Symbol Isolation*, in which the punctuation symbols are isolated are shown in Table 4.5. As observed, parsing of punctuation marks improve the system performance in both corpora.

4.3.3 Feature Extraction Methods

Hitherto we have performed experiments using a standard feature extraction method. However, selecting an appropriate feature extraction methods is crucial, as it can help to reduce the redundancy and variability of the input data. In this section, we compare three feature extraction methods. First, the previously used *Derivative-based* method. Next, the *Geometric-based* method introduced in (Bunke et al., 2004), and that has been used in state-of-the-art HTR systems (Graves et al., 2009). It must be noted that, these two feature extraction methods are implemented in the GIDOC prototype (Appendix A). Finally, the feature extraction method applied in (Dreuw et al., 2008). This feature extraction method is obtained by applying a Principal Component Analysis (PCA) (Pearson, 1901) transformation to each column and its corresponding context, which is similar to an standard ASR sliding window feature extraction. This process models a column using its whole context. However, instead of considering all the pixels within the context, it only selects the most informative ones (or a combination of them) using PCA. This feature extraction was obtained with the RWTH ASR toolkit (Rybach et al., 2009) using all data available in the database. In the following, it will be referred as *PCA window-based* feature extraction method. Table 4.6 shows the results in terms of WER in the validation set. These results corresponds to the best system obtained for each feature extraction methods, in which all the previously described parameters are tuned individually.

Table 4.6: Results in GERMANA and RODRIGO comparing different feature extraction methods

	Derivative-based	Geometric-based	PCA window-based
GERMANA	54.07	57.33	52.52
RODRIGO	45.41	54.68	39.82

As observed, the PCA window-based method achieved the best results. This is mainly caused because this feature extraction method manage to represent better the input data compared to the other methods. On one hand, the geometric-based method does not take into account the context of each pixel. On the other hand, the derivative-based method relies on the assumption that derivatives obtained from a context are informative enough. However,

a data-driven approach, such as PCA, obtains more informative features from a bigger context. Concretely, the best system uses a context of seven pixel columns. In addition, the feature vectors obtained from the PCA window-based method are smaller than the previous derivative-based features, with 50 and 60 features per vector, respectively, which results in an important computational saving.

4.3.4 Explicit blank recognition

Despite the fact that punctuation marks have been isolated, the ratio of OOVs is still quite high, specially in GERMANA, causing a high number of errors. Typically, when a HTR system encounters an OOV, it is recognise with a similar word. For instance, in GERMANA, the system recognised the words "directa" and "mente" instead of the OOV word "directamente". In the previous Section 4.3.1, we described that in the lexicon is stated that each word can be generated by its corresponding characters, and alternatively, its corresponding characters and the blank character. Concretely, in the previous example, the system recognised "directa" without an ending blank, and "mente" with an ending blank. As observed, if blanks were only considered when they are explicitly recognised, the OOV word "directamente" would have been recognised by the concatenation of words present in the lexicon. We repeated the best performing experiments but considering word splits only when the blank character is recognised. Again, all system parameters are optimised. Results are presented in Table 4.7.

Table 4.7: Results in GERMANA and RODRIGO when using the explicit blank word division.

	Baseline	Explicit Blank
GERMANA	52.52	43.70
RODRIGO	39.82	47.96

From the obtained results, it can be observed that explicit blank splits only improved the results for the GERMANA database. This is mainly caused by the fact that GERMANA possesses a higher ratio of OOVs compared to RODRIGO in their validation blocks, along with a more difficult language structure. However, this improvement has a major drawback. In those cases that an OOV is recognised is because the language model has been almost ignored. For instance, in the previously presented example, "directamente" would be recognised by recognising "directa" followed by "mente", which is indeed an nonexistent bigram in training. In fact, in RODRIGO, this method achieved worse results because the ratio of OOV is lower and it is better to rely on the language model. In conclusion, the improvement of this method is limited to those cases, in which language model estimation is poor and the ratio of OOVs is high.

4.3.5 Results on the whole document

In the previous sections, we have described the process followed to obtain the baseline system in both database, GERMANA and RODRIGO. This baseline has been obtained from the transcription of the first thousand lines of both documents, which have been used to select

methods and tune the necessary parameters to use in the transcription of the remainder. It must be noted that, the model parameters n and g, the recognition parameters, as well as the feature extraction method, remain unchanged for the rest of the experiments. In this section, we present the results of a fully supervised approach to the transcription of both databases using the best system obtained in previous experiments. It must be noted that, some pages were excluded as they contained rare document layouts, as graphics or genealogical trees. Therefore, the statistics of the databases used in these experiments are:

	GERMANA	RODRIGO
Pages	764	853
Lines	20529	20357
Running words (K)	217	232
Vocabulary size (K)	27.1	17.3
Out-Of-Vocabulary(%)	25.7	11.9
Perplexity	274.1	177.1

Table 4.8: Statistics of GERMANA and RODRIGO. Out-of-vocabulary words correspond to the percentage of running words in the test set, which do not appear in the training set. Perplexity is calculated using a ten-fold validation on the whole document.

In this approach, starting from block 3 to the last. First, the block is recognised and fully supervised to obtain its reference. Next, the quality of the recognised transcription is measured in terms of WER. Finally, the supervised block is added to the training set and the system is fully retrained from scratch. This experiment corresponds to the sequential transcription of a document using an HTR system that is continuously retrained. System is only retrained after a block is processed due to the high computational cost needed. Figure 4.5 shows the results for GERMANA while Figure 4.6 shows the results for RODRIGO.

As observed, in GERMANA the results strongly depends on the recognised block. This is mainly caused by the non homogeneous structure of the document, in which almost all blocks are not representative of the other. A further analysis revealed that errors were mainly caused by two different factors. First, half of the errors are typically caused by OOVs. OOV in GERMANA are mainly caused by the multilinguality. For instance, the first 3K lines correspond to a biography solely written in Spanish. In these blocks, the system results continuously improved from supervised blocks. However, since the 3K-th line new language, Catalan, appears with new words and a new language structures, which results in OOV, and difficulties language modelling. The interested reader is referred to (del Agua et al., 2011) for deeper analysis of the results from a multilingual point of view. Second, changes in the document structure resulted in a substantial increment of the error, due to the change in the language structure. For instance, the last two blocks correspond to the *Back matter*, which mostly contains lists and indices. Even though, these blocks do not contain a high quantity of OOV, the language model obtained from previous blocks did not correctly modelled the new language structure, which produces a high number of errors. For example, a whole chapter

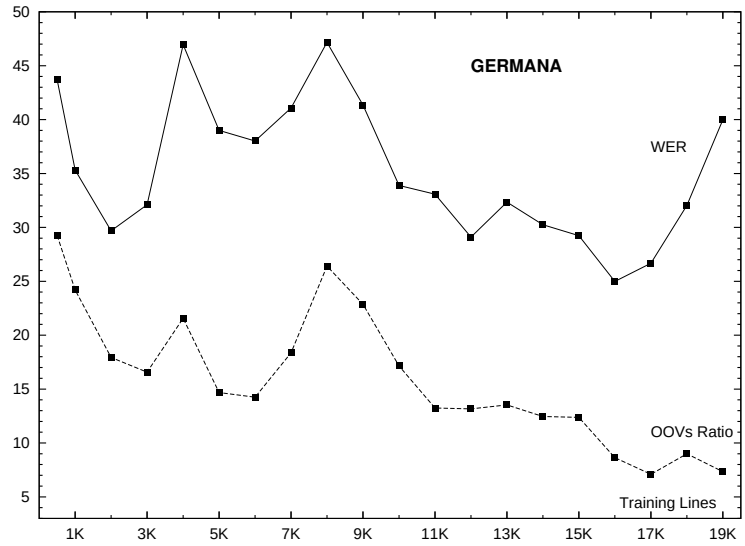


Figure 4.5: Recognition results on GERMANA for each block

corresponds to a list of important belongings to GERMANA. The lines of this chapter contain three of four words on average, and most of its words are singletons.

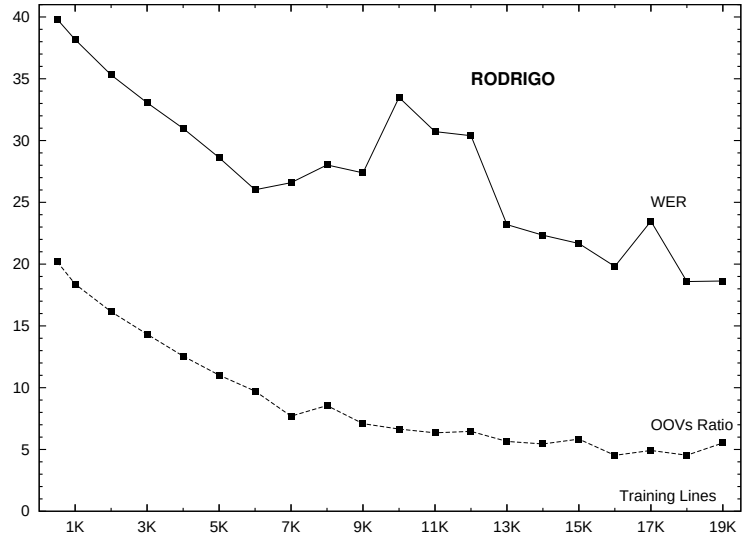


Figure 4.6: Recognition results on RODRIGO for each block

RODRIGO results are much different from GERMANA. RODRIGO is more homogeneous, and as it can be observed in Figure 4.6, almost all blocks are representative of the rest.

Results behave as expected, each time a block is added to the training set, the recognition of the next block improved. However, there is an increment of WER around line 11K. A posterior analysis revealed that at this point the author started to write words more closely, which caused that previously estimated HMMs did not correctly model the produced tighten word images. On the other hand, OOVs did not produce as much errors as in GERMANA. In fact, when half of the blocks have been supervised, the OOV ratio did not increment significantly. From this point onwards, system improvement is produced by a better character image and language modelling, each time a block is supervised.

4.3.6 Closed vocabulary recognition

Previous results showed that automatic recognition on GERMANA and RODRIGO is highly affected by the low number of data to train the language model. Concretely, the vocabulary, which in case of GERMANA accounts for half of the errors in most of the document. In order to better study this problem, we repeated the previous experiments but when lexicon is closed, i.e. there are no OOV words. In this case, image models are trained equally as before but all OOVs are added to the lexicon and the LM. It must be noted that we do not add new samples to the LM, only the OOV words. The obtained results will be unrealistic, as models are trained from data from the reference, however, the main motivation is to isolate how much accuracy is lost due to OOVs. Results for GERMANA are shown in Figure 4.7, while RODRIGO results are depicted in Figure 4.8.

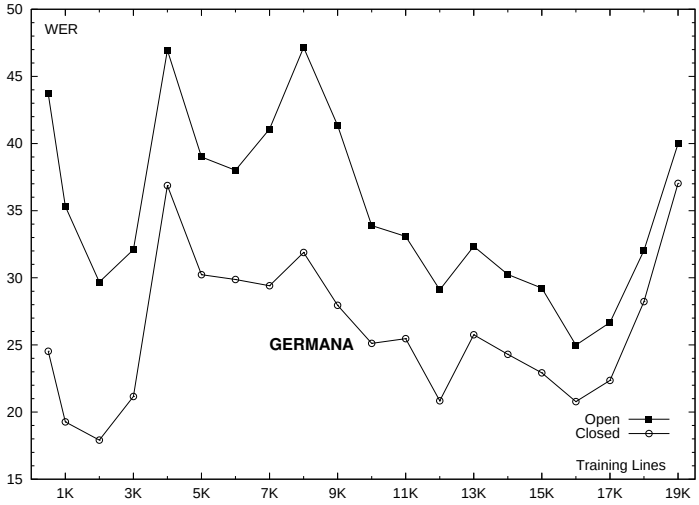


Figure 4.7: Recognition results on GERMANA for each block with closed vocabulary.

As expected, the results from closed vocabulary approach are better than then open vocabulary approach for both databases. This improvement is mainly produced by the inclusion of all OOV words, which could not be recognised in any way. In fact, the improvement is

directly proportional to the ratio of OOVs, which can be observed in Figures 4.5 and 4.6. In GERMANA, the biggest differences between the open and closed vocabulary occur in two cases. First, at the firsts blocks, where the amount of training samples is small ans so the vocabulary, and second, in blocks in which a new language appear, which introduces a high quantity of words to the vocabulary. In other cases, the differences are proportional as the remaining errors are produced by misrecognition of known words, mostly due to the image character models. On the other hand, in RODRIGO, once a sufficient quantity of data is available (around 6K lines), the difference between the open and closed approach remains static. As in GERMANA, these remaining errors are caused by the image character models.

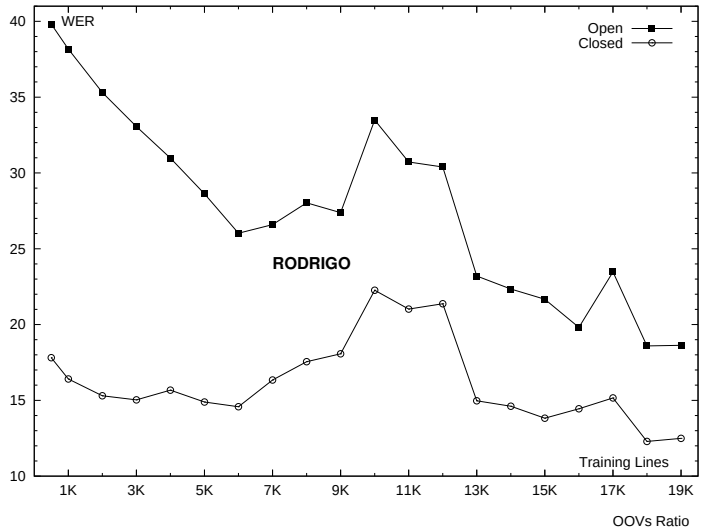


Figure 4.8: Recognition results on RODRIGO for each block with closed vocabulary.

Despite the fact that these results are unrealistic. They served to reassure that OOVs are the cause of most of the errors of the baseline system. The inclusion of these words is crucial for improving the system. This could be performed resorting to external resources, as it will be shown in the next section, or by its annotation by an external user.

4.3.7 External Resources

In the last set of experiments, we studied the inclusion of external resources in the experiments of Section 4.3.5. The main motivation is to study the possible improvement due to a better LM estimation. As seen in previous experiments, data scarcity in both corpora results in a bad estimation of the LM. In addition, the first experiments, in which few data is available, are trained from only a few lines, degrading the estimation of the LM, and the recognition results. Adequate external resources could help the system to better estimate the LM, solving the two problems: OOVs and insufficient data at the beginning of the experimental setup.

In these experiments, image models estimation remain unchanged, while the LM is now trained as a bigram LM mixture from two independent bigram models. First, an internal

(or in-domain) LM trained from all available data from the corpora, and last, an external (or out-domain) LM trained from Google n-grams (Michel et al., 2010). Google n-grams is the result of the automatic OCR of millions of scanned books, and even though Google n-grams include data from year 1534 to the present, only a few quantity of the data is from before 1800. However, some improvements could be obtained by an efficient adaptation. The optimisation of mixture parameters is performed on the first block using the EM algorithm (Iyer et al., 1994), optimising the perplexity. The resulting parameters remain unchanged for the rest of the process. On the other hand, the vocabulary is estimated from the 20k most frequent words on Google n-grams along with all words from the internal LM. Following the same framework as in Section 4.3.5, we performed the sequential transcription of GERMANA and RODRIGO. Results are presented in Figure 4.9 and Figure 4.10, for GERMANA and RODRIGO, respectively, along with the results of using only the internal models. Specifically, results using external resources are labelled as "External", while the previous results are labelled as "Internal".

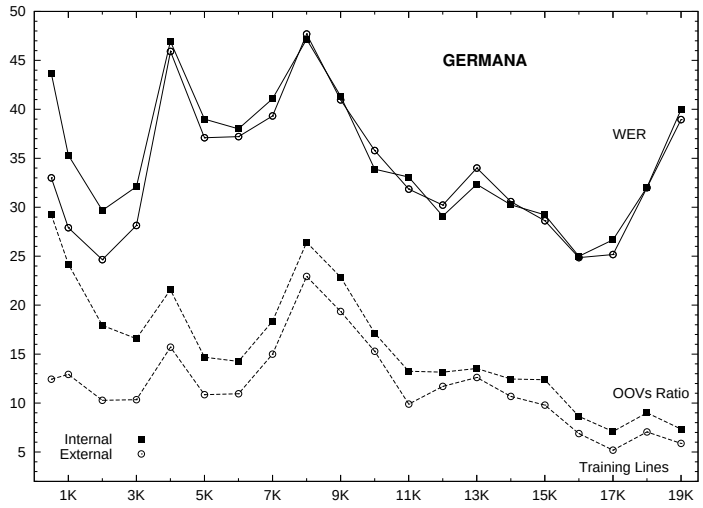


Figure 4.9: Recognition results on GERMANA for each block with closed vocabulary.

Results on GERMANA show an improvement only on the recognition of the first blocks. This is mainly caused by the high number of initial OOVs. As observed, in the first block, estimating the LM with only internal data leaded to almost 30% of OOV, while in the external case it decreased to 12.5%. Consequently, recognition rate of this block decreased almost 10 of WER. However, the improvement greatly decays once sufficient internal data is available, as amount of internal data to train the language increases. Finally, when multiliguality appears, external and internal results are practically equivalent, which is in part caused because only the Spanish part of Google n-grams was used, as it is the most predominant language in GERMANA.

Differently from GERMANA, the use of external resources in RODRIGO superseded

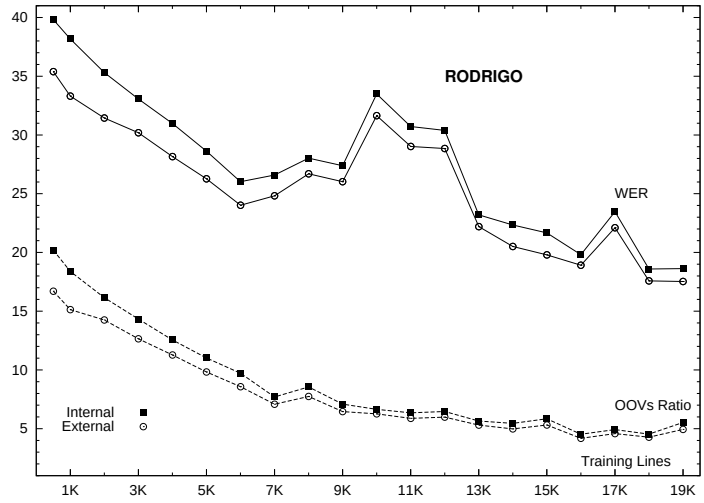


Figure 4.10: Recognition results on RODRIGO for each block with closed vocabulary.

the results from the internal approach in the recognition of all blocks. This improvement is higher at the beginning when there is not much data available, and thus, OOVs ratio reduction improved the results by 5 points of WER. Nevertheless, the improvement is reduced to 2 of WER on average when the ratio of OOVs is not further reduced. In conclusion, even though the RODRIGO corpus corresponds to an old manuscript, which vocabulary is not fully contained in the external resources used, LM adaptation helped to slightly improve the results.

4.4 Conclusions & Future Work

In this chapter, we presented two databases for handwritten text recognition and document layout analysis. We described the digitisation and annotation process that has been followed, along with a deep analysis of each document characteristics. Baseline experiments were computed to study the performance of a fully supervised approach to the document transcription. The baseline system was obtained from the optimisation of training and recognition parameters, the feature extraction method used, isolation of punctuation marks, and a word generation from explicit blanks (if needed). We also presented experiments when lexicon is closed, and using external resources to estimate the LM. The results showed that current errors are mostly produced by the language structure as well as the vocabulary of the old text documents presented. Finally, results showed that these tasks are perfectly suited for an interactive approach, as its automatic transcription is far from perfect, but within the range in which user interaction may be useful (Luz et al., 2008). It must be noted that, the presented results in this chapter are better than those previously published in each database paper (Pérez

et al., 2009; Serrano et al., 2010).

In the remainder of the thesis the HTR system will be built using the HTR system described in this section, when no external resources are employed. Even though external resources slightly improved the results, their inclusion highly difficulties the training process, as it greatly increments the size of LM and lexicon. However, isolated experiments prove that the improvement provided, external resources is similar independently from the interactive approach used.

Future work on improving the baseline system includes the application of some state-of-the-art system use discriminative systems based of recurrent neural networks (Graves et al., 2009), which could improve the recognition results. On the other hand, OOVs and language modelling problems by sub-word based recognition (Agua et al., 2012).

Preliminar versions of the work presented in this chapter has led to two publications in international conferences:

- D. Pérez, L. Tarazón, **N. Serrano**, F. Castro, O. Ramos and A. Juan. The GERMANA database. *Proceedings of the 10th International Conference on Document Analysis and Recognition (ICDAR 2009)*. Barcelona (Spain). July 2009.

- **N. Serrano**, F. Castro and A. Juan. The RODRIGO database. *Proceedings of 7th Language Resources and Evaluation Conference (LREC 2010)*. Valletta (Malta). May 2010.

Bibliography

Biblioteca de Castilla-La Mancha. http://pagina.jccm.es/biblioclm.

M. Agua, N. Serrano, J. Civera, and A. Juan. Character-based handwritten text recognition of multilingual documents. In *Proc. of Advances in Speech and Language Technologies for Iberian Languages (IBERSPEECH 2012)*, pages 187–196, 2012.

E. Belenguer, editor. *Germana de Foix, última reina de Aragón*. Universitat de València, 2007.

R. Bertolami and H. Bunke. Hidden Markov model-based ensemble methods for offline handwritten text line recognition. *Pattern Recognition*, 41:3452–3460, 2008.

BiValDi. Biblioteca Valenciana. http://bv.gva.es/.

H. Bunke, S. Bengio, and A. Vinciarelli. Offline recognition of unconstrained handwritten texts using HMMs and statistical language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(6):709 –720, 2004.

S. F. Chen and J. Goodman. An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, 13(4):359 – 393, 1999.

M. A. del Agua, N. Serrano, and A. Juan. Language identification for interactive handwriting transcription of multilingual documents. In *Proc. of the 5th Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA 2011)*, pages 596–603, Las Palmas de Gran Canaria (Spain), jun 2011.

P. Dreuw, S. Jonas, and H. Ney. Whitespace models for offline arabic handwriting recognition. In *Proc. of the 19th International Conference on Pattern Recognition (ICPR 2008)*, pages 1–4, 2008.

B. Efron and R. J. Tibshirani. *An Introduction to Bootstrap* . Chapman & Hall/CRC, 1994.

A. Giménez. Adria's kit. `http://aktoolkit.sourceforge.net`, 2011.

A. Graves, M. Liwicki, S. Fernández, R. Bertolami, H. Bunke, and J. Schmidhuber. A novel connectionist system for unconstrained handwriting recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(5):855–868, 2009.

R. Iyer, M. Ostendorf, and J. R. Rohlicek. Language modeling with sentence-level mixtures. In *Proc. of the workshop on Human Language Technology (HLT 1994)*, pages 82–87, 1994.

S. Luz, M. Masoodian, and B. Rogers. Interactive visualisation techniques for dynamic speech transcription, correction and training. In *Proc. of the 9th Int. Conf. on Human-Computer Interaction (CHINZ 2008)*, pages 9–16, Wellington, New Zealand, 2008.

U. V. Marti and H. Bunke. The IAM-database: an English sentence database for off-line handwriting recognition. *International Journal on Document Analysis and Recognition (IJDAR)*, pages 39–46, 2002.

J.-B. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, T. G. B. Team, J. P. Pickett, D. Holberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M. A. Nowak, and E. L. Aiden. Quantitative analysis of culture using millions of digitized books. *Science*, 2010.

A. Millares and J. M. Ruiz. *Tratado de paleografía española*, volume 1. Espasa-Calpe, 3rd edition, 1983.

M. Pastor. *Aportaciones al reconocimiento automático de texto manuscrito*. PhD thesis, Dep. de Sistemes Informàtics i Computació, 2007.

K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(6):559–572, 1901.

D. Pérez, L. Tarazón, N. Serrano, O. Ramos-Terrades, and A. Juan. The GERMANA database. In *Proc. of the 10th Int. Conf. on Document Analysis and Recognition (ICDAR 2009)*, pages 301–305, Barcelona (Spain), 2009.

RodrigoMCU. The RODRIGO database: digitized data. `bvpb.mcu.es`, 2006.

D. Rybach, C. Gollan, G. Heigold, B. Hoffmeister, J. Lööf, R. Schlüter, and H. Ney. The rwth aachen university open source speech recognition system. In *Proc. of the 10th Annual Conf. of the Int. Speech Communication Association (INTERSPEECH 2009)*, pages 2111–2114, Brighton, UK, 2009.

N. Serrano, F. Castro, and A. Juan. The RODRIGO database. In *Proc. of the 7th Int. Conf. on Language Resources and Evaluation (LREC 2010)*, pages 2709–2712, 2010.

A. Stolcke. SRILM - An Extensible Language Modeling Toolkit. In *Proc. of the Int. Conf. on Spoken Language Processing*, pages 901–904, 2002.

CHAPTER 5

Interactive Handwriting Recognition with limited user effort

Contents

5.1 Introduction

State-of-the-art technologies for HTR are still far from perfect both in, unconstrained domains (Bertolami and Bunke, 2008; Graves et al., 2009; Likforman-Sulem et al., 2007; Toselli et al., 2004), and in old text documents (Fischer et al., 2009). Thus, post-editing machine-generated output is not clearly better than simply ignoring it and transcribing the document from scratch. To circumvent this problem, HTR systems can be used within a CAT framework, in which both, the system is guided by the user, and the user is assisted by the system to complete the transcription task as efficiently as possible. In interactive systems, the main aim is to employ user effort efficiently since it is expensive and limited.

Interactive systems have been applied successfully to complete transcription task in many different applications, such as HTR (Toselli et al., 2007), ASR (Revuelta-Martínez et al., 2012) or syntactic tree annotation (Sánchez-Sáez et al., 2010). All these approaches reduce the quantity of user effort needed to obtain the required output, but, this quantity is not known in advance, as it depends on the number of errors on recognised transcriptions. However, in some applications, user effort may be limited because its economic or time cost. In this case, errors are expected to remain in the transcription after the whole interactive process has been carried out. Therefore, in this application, the objective of interactive systems is to obtain the best possible transcription using this limited user effort. This means that we are accepting an amount of residual error in our transcriptions in order to save user effort. For instance, an automatically transcribed document, that has been partially supervised by an user, may contain a small number of errors, and thus, it can be sufficient to convey the meaning. Similarly, there are many applications dealing with tasks that tolerate an erroneous input. For example, the output of an Automatic Speech Recognition (ASR) system can be successfully used as input in known tasks such as, dialogue act annotation (Stolcke et al., 2000), information retrieval (Grangier et al., 2003), or speech-to-speech translation (Matusov et al., 2006). All these applications may not require perfect annotation of the data, but only a sufficiently good annotation that guarantees the desired accuracy at lower user effort. In this scenario, the ideal interactive approach achieves the required transcription accuracy at the minimum user effort.

In this chapter, we describe a novel interactive approach to transcribe (old) text documents in which user effort is considered to be limited. The aim is to build a system, which employs the limited user effort to generate the best possible transcriptions as efficiently as possible following the investigations of the previous chapter. Basically, the system employs the limited effort by supervising only hypothesised words that are likely to be misrecognised (Tarazón et al., 2009). Thus, limited user effort is efficiently focused only on the supervision of the output parts for which the system is not confident enough. Low confidence words are presented to the user in isolated boxes, in a similar way as in (Ahn et al., 2008), focusing user attention and preventing them from wasting effort in reading their context. Once user supervisions has been performed, the system recomputes the transcription subjected to user supervised words by means of a constrained-Viterbi search. In this way, output errors in the unsupervised parts can be automatically amended without user supervision. At the end of the process, partially supervised transcriptions are used to improve the current system performance by means of adaptation techniques. These techniques improve the underlying system models by retraining from correctly transcribed words and high confidence parts within the transcriptions.

The remainder of this chapter is organised as the interactive process described. First, in Section 5.2, we introduce confidence measures in HTR and explain how to calculate them. Section 5.3 details how incorrectly recognised words are located by the system. Next, in Section 5.4, we explain how the system interacts with the user, and thus the type of corrections that can occur. Hypothesis recomputation constrained to user interactions is thoughtfully described in Section 5.4.1. Section 5.5 is devoted to the explanation of how the system is adapted from user interactions. Next, in Section 5.6, all the experiments performed are described and analysed. Finally, conclusions are summarised and the future work is discussed in Section 5.7.

5.2 Confidence Measures

Given a recognised word or sample, a confidence measure (CM) is score (preferably between 0 and 1), that indicates the reliability on the recognition produced by an ASR system. As described in Chapter 2, there is huge interest in computing a good CM as it is an important input in many applications, such as AL, or SL. In this thesis, we have used the CM proposed by Wessel and Ney (2005) for ASR. They proposed to directly use the posterior probability of Eq. 2.1 as a CM. The posterior probability is expected to be a good CM, as it represents the probability of the model for a sequence of words given an input image. In well estimated models, posterior probabilities of recognised words measure the uncertainty of the system on these words, and it is directly related to the correctness of its output, as not well estimated events are likely to result in errors.

However, there are two main problems in this approach. First, as said there is the segmentation problem between the input and its corresponding transcription. This problem is solved by calculating the posterior probability over a defined segment. Given an input image feature vector representation $\mathbf{x}$, and a word w from the frame s to the frame t in $\mathbf{x}$, the posterior probability can be calculated as

$$p(w \mid \mathbf{x}_s^t) = \frac{p(\mathbf{x}_s^t \mid w)p(w)}{p(\mathbf{x}_s^t)} \tag{5.1}$$

Last, in contrast to Eq. 2.1 the denominator term $p(x)$ remains because of the absence of the *argmax* operator. This denominator represents the probability of an input segment, in HTR an image segment. This probability is hard to compute as has to consider the probability of an image. A solution is to decompose it in a more intuitive way as

$$p(\mathbf{x}_s^t) = \sum_{\mathbf{w}} p(\mathbf{x}_s^t, \mathbf{w}) = \sum_{\mathbf{w}} p(\mathbf{x}_s^t \mid \mathbf{w})p(\mathbf{w}) \tag{5.2}$$

In this form, we observe that it requires the calculation over all possible word sequences. However, the probability of most of sequences is almost zero, and the summation is dominated by few ones. In consequence, we can approximate the latter probability with a smaller set of $\mathbf{w}$. This approximation will be good as long as the selected set is a good representation over all the possible sequences.

As said in Section 2.4.1,the Viterbi algorithm calculates the most probable hypothesis by efficiently exploring all possible sequences. A simple modification of this algorithm enables

us the possibility to store this set of possible sequences in form of a word graph (Wessel et al., 2001). A word graph represents, in a compact form, large sets of transcriptions. Each node in the word graph represents a time frame of $\mathbf{x}$ given a word story (in the case of bi-grams a simple word), and each arc represents the probability of generating a word from one (node) time frame to another. In this form, it is easy to compute the posterior probability of a word, as we can employ well-known graph algorithms. For each arc in the word graph, we only have to compute the ratio between the probability mass going through this edge and the probability of the whole graph. However, as reported in (Wessel and Ney, 2005), this direct posterior probability does not work well as CM. In order to better illustrate this problem, consider the example in Fig. 5.1, where a small (pruned) word graph is aligned with its corresponding text line image, and its recognised and true transcriptions are shown above and below the image, respectively.

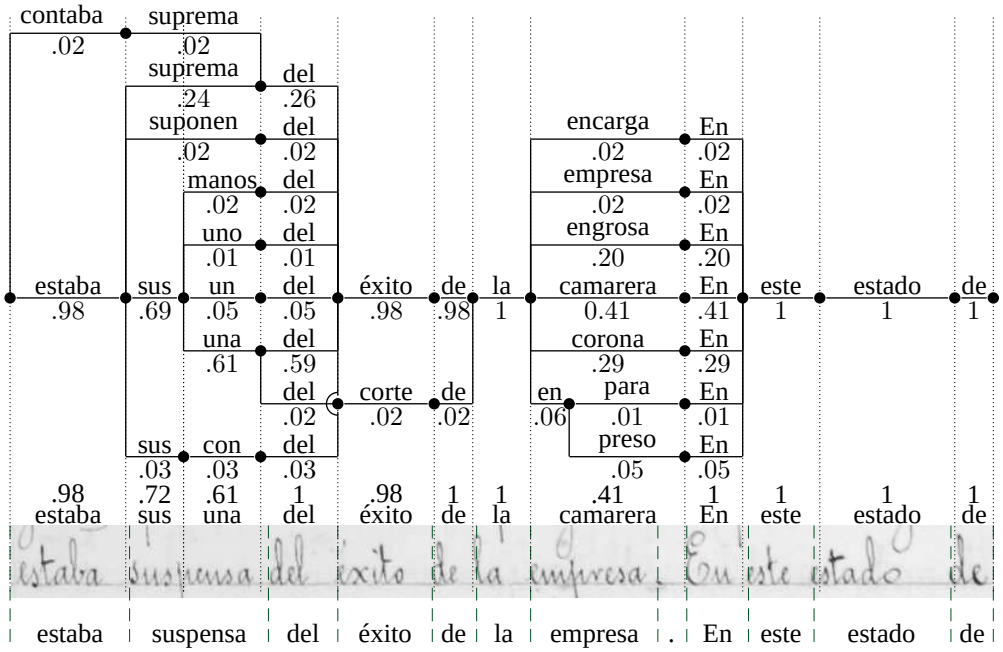


Figure 5.1: Word graph example aligned with its corresponding text line image and its recognised and true transcriptions. Each recognised word is labelled (above) with its associated confidence measure.

Each word graph node is aligned with a discrete point in space, and each arc is labelled with a word (above) and its associated posterior probability (below). For instance, in Fig. 5.1, the word "sus" has a posterior probability of 0.69 to occur between "estaba" and "un", and 0.03 to occur between "estaba" and "con". If the best hypothesis contains "estaba sus con", the word "sus" might be considered an incorrect word, as its posterior is small, while it posses a higher posterior probability in almost the same segment but for another hypothesis. This is

mainly caused because a word can be segmented in many ways even when corresponding to the same transcription, as each hypothesis segments the words differently.

In order to solve this problem, Wessel and Ney (2005) proposed to calculate the confidence measure of a word by considering all its corresponding instances in overlapping segments. Note that all word posteriors sum to one at each point in space. Therefore, the posterior probability for a word w to occur at a specific point p is given by the sum of all arcs labelled with w that are found at p; e.g. "sus" has a posterior probability of 0.72 at any point in which the two arcs labelled with "sus" are simultaneously found. Therefor, the confidence measure of a recognised word is calculated from these point-dependent posteriors, by simply summing over all points where it is most likely to occur (Viterbi-aligned). As an example, each recognised word in Fig. 5.1 is labelled (above) with its associated confidence measure.

Finally, an additional refinement is possible adding an scaling parameter called Acoustic Scale Factor (ASF) α to Eq. 5.1

$$p(\mathbf{w} \mid \mathbf{x}_s^t)^\alpha = \frac{(p(\mathbf{x}_s^t \mid \mathbf{w})p(\mathbf{w}))^\alpha}{p(\mathbf{x}_s^t)^\alpha} \tag{5.3}$$

The motivation of this parameter is to alleviate possible numerical problems due to the fact that most of the probability mass typically correspond to the best hypothesis, hence, the differences between probabilities are very small.

5.3 Active Learning: Selecting words to be supervised

The first step to efficiently use the effort of real users is to employ it in supervising incorrectly recognised words. These words typically correspond to those that the system cannot explain sufficiently, which is typically caused by their absence or scarcity in the training set. On other words, incorrectly recognised words usually correspond to those words that are not correctly estimated. Therefore, incorrect words are likely to be those words which the system is not confident enough. Active learning (AL) (Settles, 2010) is an area of ML that deals with this same problem. Concretely, it studies how to select the supervision of which recognised samples will improve the current system the most. One of the most widespread and straightforward methods of AL is called *uncertainty sampling*, which selects which samples are supervised in terms of their posterior probability.

As defined in the previous section, the posterior probability of a recognised word can be used as a CM. In this case, the best way to detect most incorrect words is to order them by its confidence measure from lowest to highest, and then, supervise them sequentially to improve the most both, the system and the resulting transcription. In order to assess the correctness of this CM to detect incorrect words, we have conducted an experiment on each validation set defined in Section 4.3 for GERMANA and RODRIGO. For each recognised word of the validation set we calculated its CM. Then, we compared three different approaches to select which words are supervised. First, *Random* selection of words, which is considered the baseline. Second, the least *Confidence* selection previously described. Last, a selection performed by an *Oracle*, which first corrects all the incorrect words. Random and oracle selections represent, the worst and best selection case approach possible, respectively. Results are shown in Figure 5.2. On one hand, the x axis indicates the quantity of words that are

supervised. On the other hand, the y axis corresponds to the percentage (over the total) of incorrect words that are detected.

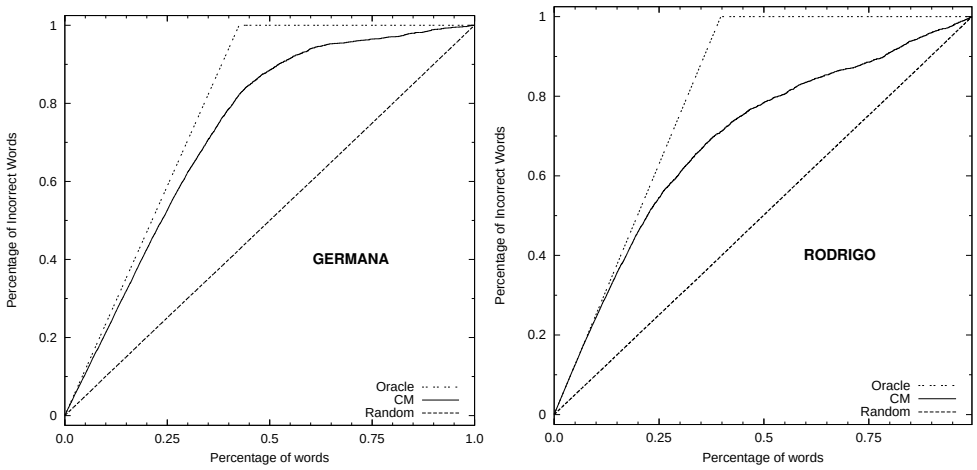


Figure 5.2: Percentage of incorrect words detected depending on the selection method.

Results clearly show that the posterior probability can be used to detect incorrect words, as it supersedes the results of the baseline random approach. As expected, words with a small CM are likely to be incorrect, while words with a high CM are typically correct. However, the results obtained using CMs are far from perfect. In both corpora, CMs only managed to detect (almost perfectly) a 20% of incorrect words, which is as good as the best approach. In fact the words are those with the least confidence. From this point onwards, CMs almost detect incorrect words randomly. In conclusion, CMs are an effective way to detect incorrectly recognised words, but its performance strongly decays once a certain percentage of the least confident are supervised.

5.4 User Supervision

As mentioned, we deal with the interactive transcription of (old) text documents in which user effort is limited. In our proposed approach, user effort is employed in supervising low confidence hypothesised words. For the sake of clarity, we detail the supervision of a recognised word from the user point of view. Figure 5.3 shows the transcription dialog of GIDOC, which is a set of tools that implements the proposed interactive transcription approach (a whole description can be found in Appendix A). In this figure, it can be observed a text line image, whose baseline is underlined in blue, has been automatically recognised and the obtained transcription is presented in line number eight. In this moment, the system asks the user to supervise a recognised word, which may be possibly incorrect. The word to be supervised is highlighted both in the image by darkening all but the corresponding word, and in the editable line by selecting it. It must be noted that word highlighting helps to focus user attention and

prevents him from reading the context whenever unnecessary, saving user effort. In this case, the recognised word to be supervised is “entonces” instead of the correct “teutonico”, which can be corrected without looking at the context. The user will simply input the correct word and move to the next supervision.

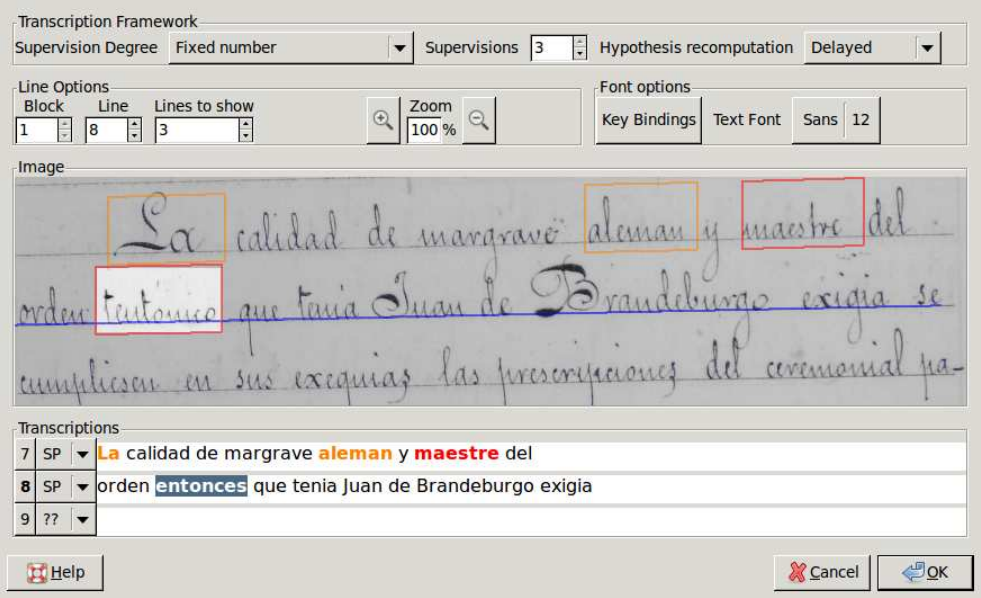


Figure 5.3: Interactive transcription of the recognised word “entonces” using GIDOC. The corresponding reference word “teutonico” is highlighted by darkening the rest.

It must be noted that the snapshot shown in Figure 5.3 is a simple user supervision. In practice, it might be the case that image segmentation and recognised word alignment are not perfect. For this reason, we need to consider the following four supervision cases:

1) The text line image segment contains a word that has been correctly recognised.

2) The text line image segment contains a word that has been incorrectly recognised.

3) The text line image segment contains more than one word.

4) The text line image segment corresponds to a portion of a word.

The first two cases simply ask the user to supervise the content of a correctly segmented word, which corresponds to the case detailed in Figure 5.3. In this situation, the user simply amends or accept the recognised word depending whether it has been misrecognised or not. An example of the third case is shown in Figure 5.1, where the supervision of the recognised word “camarera” would result in two user edition operations: the substitution of this word by “empresa” and the insertion of “.”. Lastly, an example of the fourth case occurs when supervising the word “una” in the same figure. In this case, the image segment cannot be

correctly identified as a single word, and consequently, the user would delete the current hypothesised word "una". Later on, if the user is asked to supervise the preceding or next image segment corresponding to a previously deleted word, such as "sus" in the figure, the system would show to the user the image segment associated with the word "sus" plus the deleted word "una", as they could correspond to a whole word "suspensa".

5.4.1 Constrained Viterbi-based search

Hitherto we have described the steps needed to locate and supervise (possibly) incorrect words. As said, transcriptions are obtained by searching the most probable hypothesis among all the possible ones. Accordingly, recognised words within the same line depend on each other, and thus, incorrectly recognised words affect their surroundings. Once some words have been supervised, a better strategy would be to modify the current system hypothesis to include them, improving the remainder.

A first approach using this idea for CAT of text images was proposed in (Toselli et al., 2007), which followed previous ideas applied to machine translation and speech recognition (Barrachina et al., 2009; Rodríguez et al., 2010). In this work the authors proposed a prefix-based interactive-predictive approach in which the user reads from left to right both, the corresponding text imagen and the system output, correcting the first incorrect word. Then, a valid prefix $\mathbf{p}$ is defined including all words up to the one corrected. Next, the system recomputes its hypothesis constrained to this (fully supervised) prefix, which may improve the unsupervised words. This process continues until all words have been supervised.

This supervision protocol updates the current hypothesis by searching for the most probable suffix $\hat{\mathbf{s}}$ that better completes the validated prefix $\mathbf{p}$. This is achieved by conveniently introducing the prefix dependency on Eq. 2.1

$$\hat{\mathbf{s}} = \arg\max_{\mathbf{s}} p(\mathbf{s} \mid \mathbf{x}, \mathbf{p}) = \arg\max_{\mathbf{s}} p(\mathbf{x} \mid \mathbf{s}, \mathbf{p})\, p(\mathbf{s} \mid \mathbf{p}) \tag{5.4}$$

In order to perform this search, the sequence of feature vectors is split into two fragments x_1^b and x_{b+1}^T, which depends only on $\mathbf{p}$ and $\mathbf{s}$, respectively. The boundary b is unknown, and considered a hidden variable, the estimation of which is approximated in the search process

$$\begin{aligned} \hat{\mathbf{s}} &\approx \arg\max_{\mathbf{s}} \sum_{1 \le b \le T} p(x_1^b \mid \mathbf{p})\, p(x_{b+1}^T \mid \mathbf{s})\, p(\mathbf{s} \mid \mathbf{p}) \\ &\approx \arg\max_{\mathbf{s}} \max_{b} p(x_1^b \mid \mathbf{p})\, p(x_{b+1}^T \mid \mathbf{s})\, p(\mathbf{s} \mid \mathbf{p}) \end{aligned} \tag{5.5}$$

This two-step interactive-predictive search defined in Eq. 5.5 is repeated until the transcription has been completely validated. As a result, error-free transcriptions are obtained.

However, the prefix-based approach presents three main limitations in our framework. Firstly, the user needs to supervise all recognised words. Thus, this approach is not applicable when user effort is limited. Secondly, supervision must be performed from left to right, and an important user effort has to be devoted to locate output errors. In order to overcome this drawbacks, we have migrated from a lattice-based search (Toselli et al., 2007) to constrained Viterbi-based search (Kristjannson et al., 2004).

As we already pointed out, the easiest way to improve the system transcription is to simply ask the user to supervise some (hopefully misrecognised) words. This simple strategy will be referred to from here on as *conventional*, and considered to be the interactive baseline system with respect to the other interactive approaches. However, user supervisions can be used to further improve the transcription beyond basic correcting. Following this idea, we proposed an extension to the conventional approach, in which given the supervision of an image segment, the system recomputes a new transcription subject to user supervisions (Serrano et al., 2010). As said, this approach has also been followed by Toselli et al. (2007), but as observed in Eq. 5.4, it is constrained to a left-to-right supervision protocol. On the contrary, in our approach any word can be supervised independently from their context. This is due to the migration from lattice-based search (Toselli et al., 2007) to constrained Viterbi-based search (Kristjannson et al., 2004). The constrained Viterbi-search allows for the definition of words that must be necessarily recognised for a given image segment during the search process. These words narrow the expansion of the search trellis at them, reducing the amount of hypothesis that are explored.

In (Serrano et al., 2010), the user performs the supervision according to the first three supervision cases previously described. As a result, the user defines a constraint $\mathbf{c} = (c_1, c_2, c_3)$ by which a word c_3 must be recognised from segment $x_{c_1}^{c_2}$ of the text line image. This constraint can be included in the general search problem (Eq. 2.1) as follows:

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} p(\mathbf{w} \mid \mathbf{x}, \mathbf{c}) = \arg\max_{\mathbf{w}} p(\mathbf{x} \mid \mathbf{w}, \mathbf{c})\, p(\mathbf{w}) \tag{5.6}$$

where the language model $p(\mathbf{w})$ is assumed to be independent of the user constraint $\mathbf{c}$. At this point, it is convenient to split the image model in accordance with $\mathbf{c}$:

$$p(\mathbf{x} \mid \mathbf{w}, \mathbf{c}) = p(x_1^{c_1-1} \mid w_1^{s-1})\, p(x_{c_1}^{c_2} \mid w_s, \mathbf{c})\, p(x_{c_2+1}^{T} \mid w_{s+1}^{|w|}) \tag{5.7}$$

where $p(x_{c_1}^{c_2} \mid w_s, \mathbf{c})$ is the only part of the image model in which the constraint $\mathbf{c} = (c_1, c_2, c_3)$ takes effect. As c_3 is the only word that can be recognised from the image segment $x_{c_1}^{c_2}$, $p(x_{c_1}^{c_2} \mid w_s, \mathbf{c})$ will be computed as:

$$p(x_{c_1}^{c_2} \mid w_s, \mathbf{c}) = \begin{cases} p(x_{c_1}^{c_2} \mid w_s) & c_3 = w_s \quad (5.8) \\ 0 & c_3 \neq w_s \quad (5.9) \end{cases}$$

for each hypothesis $\mathbf{w}$ and any position s in which w_s is to be considered as the word written by hand in the image segment $x_{c_1}^{c_2}$. On the other hand, the image models for the prefix and suffix, $p(x_1^{c_1-1} \mid w_1^{s-1})$ and $p(x_{c_2+1}^{T} \mid w_{s+1}^{|w|})$, are assumed to only depend on the given word sequences.

As a novelty, we further extend in this work the approach presented in (Serrano et al., 2010) to include the supervision of words that need to be deleted (Serrano et al., 2013), i.e. the fourth supervision case described above (e.g. deletion of "sus" or "una" in Figure 5.1). Now, the user defines a constraint $\mathbf{c} = (c_1, c_2, \bar{c}_3)$ by which word c_3 should not appear in any segment $\left(x_{k_1}^{k_2}\right)$, totally or partially, within segment $x_{c_1}^{c_2}$. Formally, Eqs. 5.7-5.9 can be extended to include the four supervision cases as follows:

$$p(\mathbf{x} \mid \mathbf{w}, \mathbf{c}) = \max_{0<k_1<k_2<T+1} p(x_1^{k_1-1}, x_{k_1}^{k_2}, x_{k_2+1}^{T} \mid \mathbf{w}, \mathbf{c}) \tag{5.10}$$

where

$$p(x_1^{k_1-1}, x_{k_1}^{k_2}, x_{k_2+1}^{T} \mid \mathbf{w}, \mathbf{c}) = p(x_1^{k_1-1} \mid w_1^{s-1})\, p(x_{k_1}^{k_2} \mid w_s, \mathbf{c})\, p(x_{k_2+1}^{T} \mid w_{s+1}^{|w|}) \quad (5.11)$$

with

$$p(x_{k_1}^{k_2} \mid w_s, \mathbf{c}) = \begin{cases} p(x_{k_1}^{k_2} \mid w_s) & \substack{[k_1,k_2]=[c_1,c_2] \\ c_3=w_s} \quad (5.12) \\ 0 & \substack{[k_1,k_2]=[c_1,c_2] \\ c_3\neq w_s} \quad (5.13) \\ 0 & \substack{[k_1,k_2]\cap[c_1,c_2]\neq\emptyset \\ c_3=w_s} \quad (5.14) \\ p(x_{k_1}^{k_2} \mid w_s) & \text{otherwise} \quad (5.15) \end{cases}$$

Note that Eq. 5.10 reduces to Eq. 5.7 when $[k_1, k_2] = [c_1, c_2]$ and, in this case, Eqs. 5.12-5.13 equal to Eqs. 5.8-5.9. The new deletion case is covered in Eqs. 5.14 and 5.15.

As explained above, constrained search generates a new hypothesis subject to user supervisions. However, as the user may ask for more than one supervision per text line image, the system could consider at least two alternative strategies regarding when a new hypothesis is recomputed . The first strategy, known as *delayed*, consists in recomputing the most probable hypothesis after all supervisions are done. To put it formally, let us assume that M constraints $\{\mathbf{c}^{(m)}\}$ ($m = 1, \ldots, M$) must be satisfied for each hypothesis $\mathbf{w}$ and positions $\{s^{(m)}\}$ (with $s^{(1)} < \cdots < s^{(M)}$) in which their corresponding words $w_s^{(m)}$ are considered to be written by hand in segments $\{(k_1^{(m)}, k_2^{(m)})\}$ (with $0 < k_1^{(1)} < k_2^{(1)} < \cdots < k_2^{(M)} < T + 1$). Then, our single-constraint model in Eq, 5.10 can be extended to multiple constraints as follows:

$$p(\mathbf{x} \mid \mathbf{w}, \{\mathbf{c}^{(m)}\}) = \max_{\{(k_1^{(m)}, k_2^{(m)})\}} p(x_1^{k_1{}^{(1)}-1} \mid w_1^{s^{(1)}-1}) p(x_{k_1{}^{(1)}}^{T} \mid w_{s^{(1)}}^{|w|}, \{\mathbf{c}^{(m)}\}) \quad (5.16)$$

with

$$p(x_{k_1{}^{(1)}}^{T} \mid w_{s^{(1)}}^{|w|}, \{\mathbf{c}^{(m)}\}) = \prod_{m=1}^{M} p(x_{k_1{}^{(m)}}^{k_2{}^{(m)}} \mid w_{s^{(m)}}, \mathbf{c}^{(m)}) p(x_{k_2{}^{(m)}+1}^{k_1^{(m+1)}-1} \mid w_{s^{(m)}+1}^{s^{(m+1)}-1}) \quad (5.17)$$

where each constraint-conditioned model $p(x_{k_1{}^{(m)}}^{k_2{}^{(m)}} \mid w_{s^{(m)}}, \mathbf{c}^{(m)})$ is computed as in the single-constraint case (Eqs. 5.12–5.15). In Eq. 5.17, it is also assumed that $k_1^{(M+1)} - 1 = T$ and $s^{(M+1)} - 1 = |w|$ (corresponding to the final image segment).

Recomputation strategies

As explained above, the constrained search generates a new hypothesis subject to user supervisions. However, as the user may ask for more than one supervision per text line image, the system could consider at least two alternative strategies regarding when a new hypothesis is recomputed. The first strategy, known as *delayed*, recomputes its most probable hypothesis after all supervisions are performed. The second strategy, referred to as *iterative*, recomputes a new hypothesis after each user supervision is committed.

Figure 5.4 shows an example of the described constrained-Viterbi on a line of GERMANA, in which the recomputation is performed after all user interactions have bee performed, i.e. "delayed" strategy. At the top of the figure the line image is shown aligned

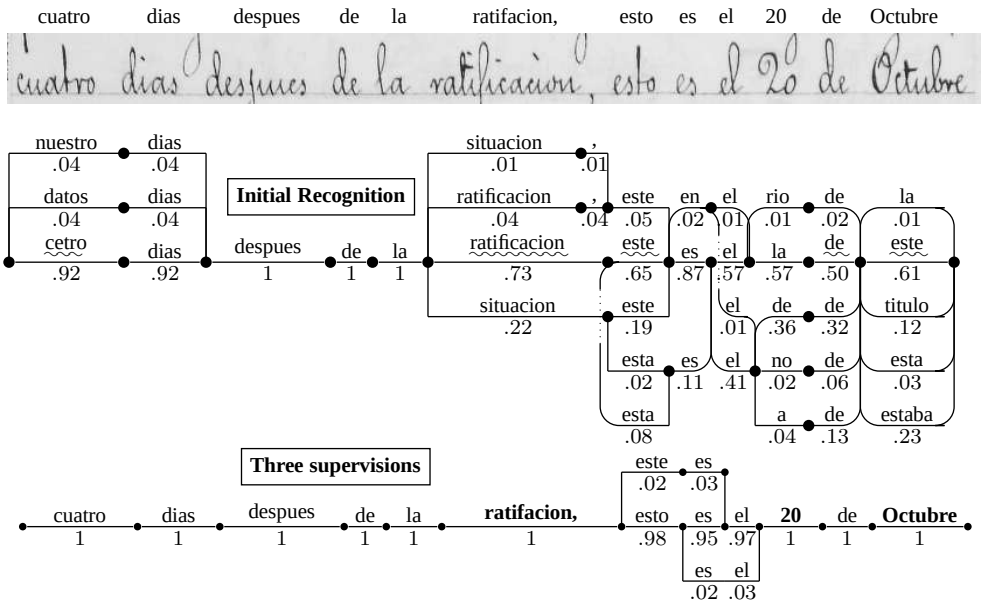


Figure 5.4: Example of *delayed* strategy in which three words are supervised. At the top, the reference text line is aligned with the image line. Just below, the initial word graph from recognition with words scored with their confidence is shown. The central row in the word graph contains the most probable hypothesis, where incorrect words are marked using a wavy line. At each iteration the user supervises the least confident word, and the system recomputes its most probable constrained hypothesis generating a new word graph.

with its transcription. Next, a pruned version of the hypothesis word-graph generated by the "Initial Recognition" is shown. The best hypothesis is shown at the middle part of the graph, in which the incorrect words are marked with a wavy line. Each arc shows a word along with its corresponding confidence measures, which are obtained as described in Section 5.2. Finally, at the bottom part of the figure, the word-graph obtained after the application of the constrained-Viterbi recomputation once three words were supervised, is presented.

As observed, user supervision of the least confident word and the posterior hypothesis recomputation, reduce the size of the word-graph, and thus, its uncertainty. In fact, not only the uncertainty of supervised segments is reduced (or even removed) but in other segments within the line image. For instance, the supervision of the middle word "ratificacion", and end words "20" and "Octubre", reduced the uncertainty of the system on the constrained-Viterbi search, and it manage to update previous recognised first word "cetro", with the correct word, "cuatro".

Alternatively, Figure 5.5 shows the result of the constrained Viterbi recomputation for the iterative strategy. In this case, similarly to the delayed figure, each supervision is shown along it corresponding word-graph. Uncertainty reduction can be better observed in this case. Concretely, we can observe that each time a supervision of a segment is committed, the uncertainty of both the surrounding segment and the rest of them is reduced. It must be

noted that recognised words that are supervised differ from the one of the delayed example. This is produced by the hypothesis recomputation performed after each supervision, which generates a whole new transcription with different confidence measures. This approach is expected to perform better because recomputation is performed continuously. However, as we see in this practical example, in contrast to the delayed strategy, an error remains at the end of the process.

For the sake of clarity, a summarised version of the previous examples from Figures 5.4 and 5.5 are presented in Figure 5.6. In this figure, the three recomputation methods are presented altogether. For each strategy and each step (if more than one), the most probable hypothesis is presented at top, followed by the following most probable ones. Supervised words are highlighted in bold, and incorrectly recognised words are marked with a wavy underline. Summarising, it can be observed that from the five original incorrectly recognised words, after three user supervisions; the conventional strategy manages to correct three errores, while the iterative strategy corrects four, and the delayed strategy of all them.

5.5 Adaptation from Partially Supervised Words

Up to this point, we have described how to select possibly incorrect recognised words, supervise them, and use this supervision to improve the system hypothesis. At the end of this procedure, the obtained transcription is constituted by supervised and unsupervised words. This transcription cannot be further amended, but, it can be used to improve the current system estimation. Consequently improving the recognition of next transcriptions. Supervised words within the transcription can be directly added as new training data, as they correspond to new samples. However, unsupervised words cannot be added right away, as its direct addition to the training data may harm the system estimation. A better idea will be to intelligently select which unsupervised words improve the system the most among of all unsupervised ones. This exact problem has been studied thoughtfully in ML by a class of learning techniques referred as Semi-Supervised Learning (SSL) (Zhu, 2006).

SSL studies how to best improve a system from unsupervised input data. One of the simplest and most successful techniques is to consider the problem as a classification problem, in which unsupervised words are classified as correct or incorrect, and then, add the correct ones to the training set. In this thesis, we have employed this approximation by following these steps. First, input data is classified in its most probable class. Next, a confidence measure is computed for this labelling. Finally, words are added to the training set if they meet a certain threshold, as they are considered correct. In our case, the most probable transcription has already been generated along with its corresponding confidence measures before user supervision. All that remains is to select an appropriate confidence threshold. In conclusion, the system performance gain depends on both, the confidence measure and the threshold defined. However, it is not straightforward how to select them.

As explained, classification is performed by selecting a threshold. All words below the threshold will be considered incorrect, and on the contrary, all words over the threshold will be considered correct. This selection produces two types of different errors: false positives (FP), which are incorrectly recognised words that are considered correct, and false negatives (FN), which are correctly recognised words that have been considered incorrect. The problem

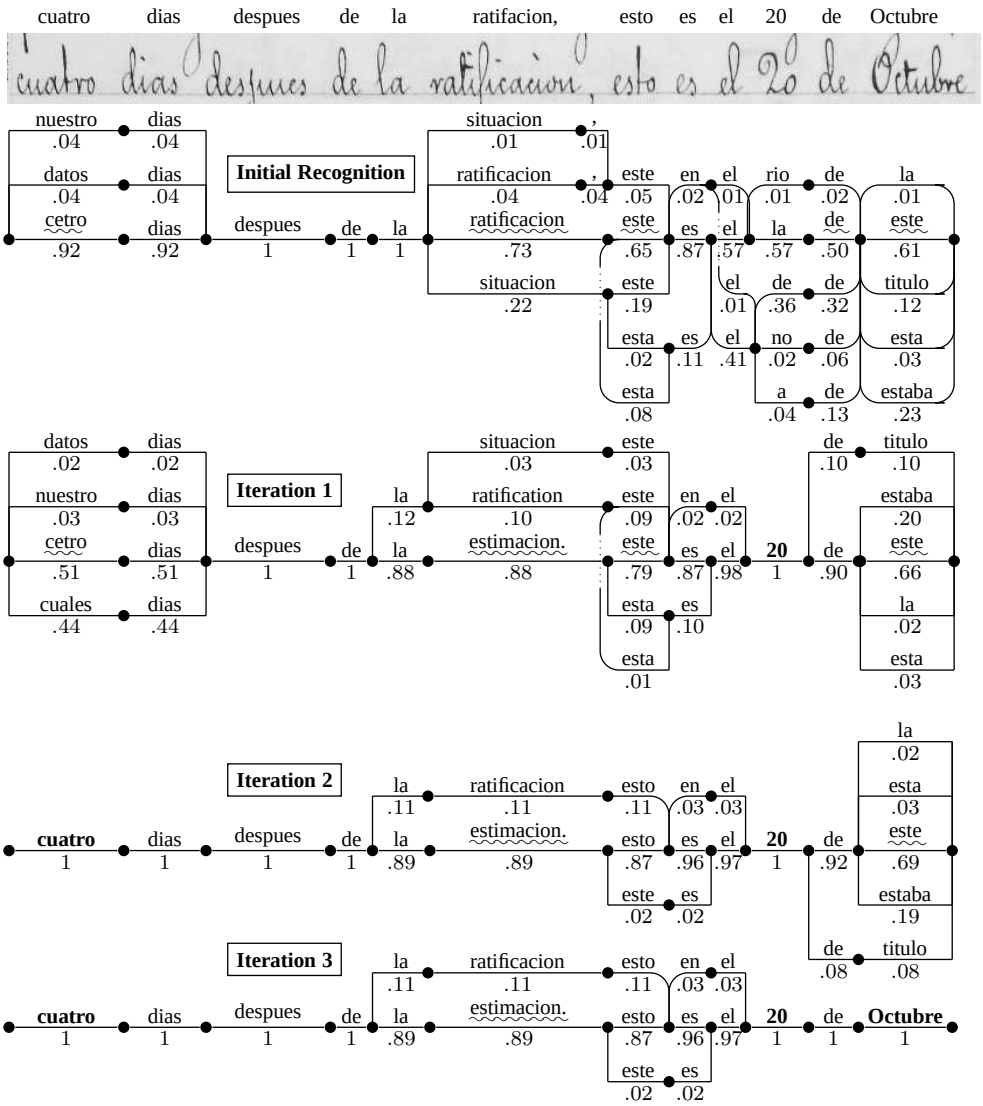


Figure 5.5: Example of *iterative* strategy in which three words are supervised. At the top, the reference text line is aligned with the image line. Just below, the initial word graph from recognition with words scored with their confidence is shown. The central row in the word graph contains the most probable hypothesis, where incorrect words are marked using a wavy line. At each iteration the user supervises the least confident word, and the system recomputes its most probable constrained hypothesis generating a new word graph.

is that, depending on the task, one type of error could be more important than the other. Thus, the error incurred from the selection of a threshold has to be defined as a combination of them.

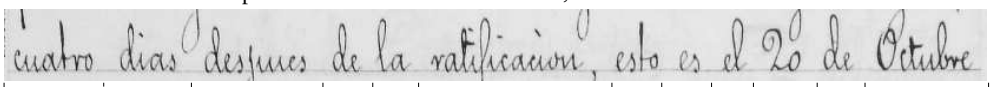
cuatro	dias	despues	de	la	ratificacion,	esto	es	el	20	de	Octubre
											
Initial											
cetro .92 nuestro .04 datos .02	dias .92	despues 1	de 1	la 1	ratificacion .73 situación .22 ratificación .04 situación .01	este .90 esta .10 , .04 , .01	es .98 es .02	el 1	la .36 de .57 a .04 no .02 rio .01	de 1	este .61 estaba .23 titulo .12 esta .03 la .01
Conventional											
cetro .92 nuestro .04 datos .02	dias .92	despues 1	de 1	la 1	**ratificacion,** 1	este .90 esta .10	es .98	el 1	**20** 1	de 1	**Octubre** 1
Delayed											
cuatro 1	dias 1	despues 1	de 1	la 1	**ratificacion,** 1	esto .98 este .02	es 1	el 1	**20** 1	de 1	**Octubre** 1
Iterative											
cetro .50 cuales .45 nuestro .03 datos .02	dias 1	despues 1	de 1	la 1	estimacion .88 ratificación .10 situación .03	este .90 esta .10	es .98 en .02	el 1	**20** 1	de 1	este .66 estaba .20 titulo .10 esta .03 la .02
cuatro 1	dias 1	despues 1	de 1	la 1	estimacion .89 ratificación .11	esto .98 esta .11	es .98 en .03	el 1	**20** 1	de 1	este .69 estaba .19 titulo .08 esta .03 la .02
cuatro 1	dias 1	despues 1	de 1	la 1	estimacion .89 ratificación .11	esto .98 este .02	es .98 en .03	el 1	**20** 1	de 1	**Octubre** 1

Figure 5.6: Comparative of the conventional, delayed, and iterative strategies when supervising a given recognised sentence. At the top, the reference is aligned with its corresponding text line image. The initial hypothesis is displayed after the image, in which each word is accompanied by its confidence. Misrecognised words are underlined using a wavy line, and alternative hypotheses for each word are shown in grayscale. The most probable hypotheses after user supervision of three words for the presented strategies are shown. The three supervised words are highlighted in bold face.

The simplest error metric is to count the number of errors, independently from its type, that has been committed among all recognised words when the classification was performed with a specific threshold. In this case, the error corresponds to the Classification Error Rate (CER). Fig. 5.7 shows the results in terms of CER when classifying the recognised words on the validation set of RODRIGO. Each curve represent a different value for the tuning parameter α of Eq. 5.2 in the confidence measures calculation, and each point of the curve represents the CER of a confidence measure threshold. In terms of CER, the lowest CER, 21.2%, is achieved using an ASF of 20 and a threshold of 0.997602. However, we can observed in the zoom-in of Figure 5.7, that the behaviour of confidence measures around the best threshold is rather unstable. This is mainly caused by the confidence measures calculation, in which most words result in a value of 1, and most of the remainder are centered around 0.95.

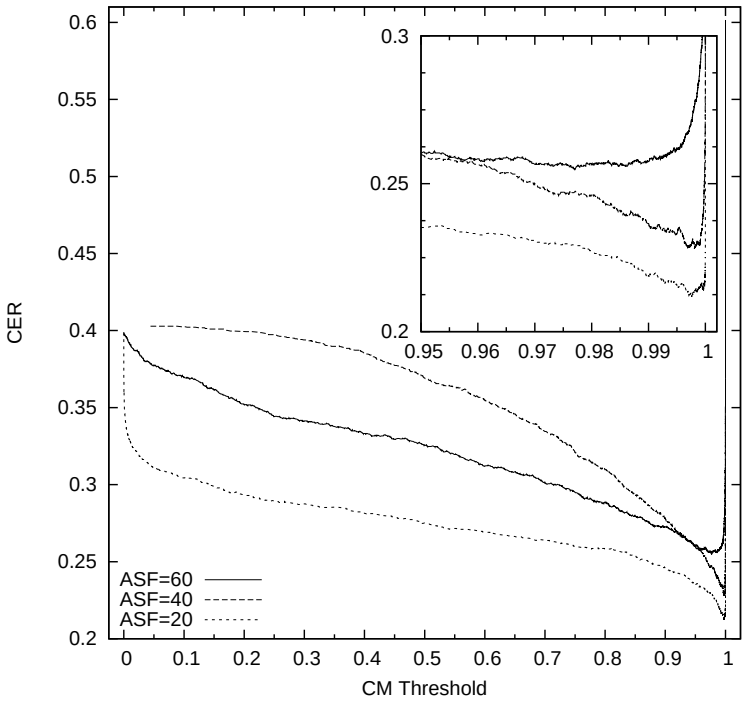


Figure 5.7: Example of CER curves when optimising the confidence measures.

A more refined metric is to represent all together the ratio of the two types of errors when varying the threshold. Concretely, for each possible threshold, the ratio of false positives and false negatives is represented in a two dimensional plot. The resulting plot corresponds to the so-called ROC curve. The study of this plot is very interesting as it shows how the two type of errors behave. Figure 5.8 shows the ROC curve for the same confidence measures used in the previous figure, Figure 5.7. Again, each curve represent a different set of confidence measures for different parameters of ASF, and each point represents the two types of errors

for a different threshold. The idea of this curve is to select a certain threshold according to a desired behaviour. For instance, in this case, if a low number of FP is desired, the number of FN will be high, and viceversa. Similarly, the best curve will be the one passing as close to the left and upper edges as possible, because it would correspond to the lowest number of errors of both types.

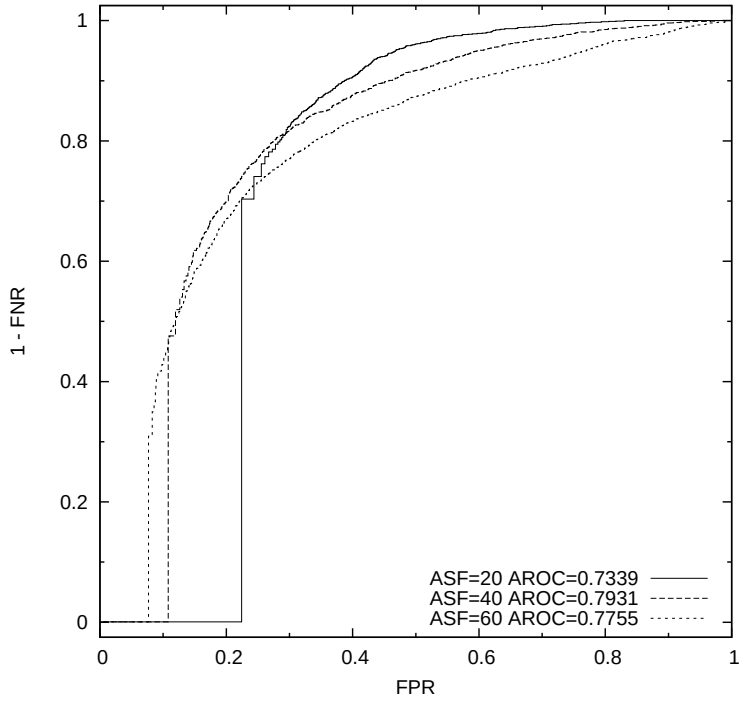


Figure 5.8: Example of ROC curves when optimising the confidence measures.

In order to compare how the curves are related to this best case, we can calculate the Area Under the ROC Curve (AROC). The closer this value is to 1, the closer the curve is to the edges. AROC values for each curve are also presented in Fig. 5.8. In terms of AROC, the best curve is obtained when using an ASF of 40. This is mainly caused because the previously commented effect, that small values of ASF cause confidence measures to be more unstable and centered around the value 1.0. An example of this behaviour can be observed for the ASF=20 curve, in which the lowest confidence measure obtains a FPR of 0.21 and FNR of 0.3. These values lead to a pessimistic calculation of AROC, as a great part of its area is missing.

As said, the objective of semi-supervised learning is to improve the recognition of the following image lines. This consideration would imply to compute for every confidence measure, and each possible threshold, the performance of recognising a validation set. However, this procedure is unfeasible as it is very time consuming. In this thesis, we have rather tested a few best values for each of the presented error metrics, CER and AROC. Experiments on

the validation set showed the CER value providing the best recognition rate of the following image lines.

Once some words have been supervised, and some of the high confidence unsupervised words have been classified as correct, there still remains one open problem, how to update the system models, HMMs and LM, from these words. As said, HMMs are trained from input images and their corresponding words, while LMs are directly trained from the transcription. Once some words within a line have been supervised, and the remaining unsupervised words are classified into correct or incorrect, the line is built from correct and incorrect segments of words. In our proposed approach, correct segments can be added directly as new n-grams for the LM estimation. However, correct segments cannot be employed directly to estimate directly the HMMs, as their segmentation on the image in unknown. Following the work of Wessel and Ney (2005), a forced Viterbi alignment can be computed to segment the image into segments. Even though, incorrect segments may also be segmented, in practice, correct segments are obtained successfully, as the number and size of words is similar to the reference.

In conclusion, supervised and high confidence unsupervised words are incorporated as new fresh training data to improve system performance, thanks to the combination of active and semi-supervised learning. We successfully adopted and tested this approach in (Serrano et al., 2009), corroborating previous results in the area of speech recognition (Hakkani-Tür et al., 2006). It must be noted that, to our knowledge, this is the first work that combines active and semisupervised learning at the word level in HTR.

5.6 Experiments

In this section, we present the experiments that have been carried out on the two presented datasets in Chapter 4: GERMANA (Sec. 4.2.1) and RODRIGO (Sec. 4.2.2) using the same partition as in Sec. 4.3.5 and a similar sequential setup. Figures in Table 4.8 reflect that GERMANA is more complex than RODRIGO, as it was shown in Chapter 4. The vocabulary size and the number of out-of-vocabulary (OOV) words are larger in GERMANA. OOV words constitute a major source of errors since they represent the percentage of running words in the test set that do not appear in the training set. Moreover, GERMANA also has greater perplexity which can be considered as the average number of words which can follow any word sequence. Note that language model perplexity is typically used to evaluate the difficulty of the task. Perplexity is calculated using a ten-fold validation on the whole document. This difference between the perplexity of both documents is due to the multilingual nature and document layout variability in GERMANA.

We simulated the interactive transcription of these two handwritten text documents using the presented approach. Due to their sequential book structure, the task is to transcribe them from the beginning to the end of the whole document. Each database was divided into 7 consecutive blocks of 3200 lines, except for the first block, which only contains 1000 lines, and the last block, which also includes the last remnant of the lines. The experimental setting for each database is as follows. The first block is devoted to train an initial system, and tune the preprocessing and recognition parameters. These optimised parameters remain the same for the rest of the experiments. Next, starting from block two to the last block, each new

block is recognised and evaluated in terms of Word Error Rate (WER). Next, the recognised block is processed to select new candidate training segments (if necessary), and lastly, added to the training set. Finally, the system is fully re-trained each time a new block is added to the training set. It must be noted that block division is performed because the complete adaptation of the models cannot be performed in real time since it takes several days in a single core.

In the remainder of the section, first, we present a user supervision model to assess our interactive HTR system. Finally, experimental results are reported in Section 5.6.2.

5.6.1 User Interaction Model

In order to evaluate the actual performance of the interactive HTR system proposed, we should carry out an evaluation campaign with real users. However, human evaluation is an expensive and time-consuming task. Alternatively, an automatic evaluation allows us to rapidly assess and compare different interactive strategies at very low cost. To this purpose, a user interaction model is defined to simulate the interaction of a real user with our interactive HTR system.

As said, we consider an interaction model in which the user is asked to supervise n recognised words of each image line in increasing confidence order. User simulation is carried out by a simple yet realistic user interaction model, which simulates the user edit operations described above. First, we compute a minimum edit (Levenshtein (Levenshtein, 1966)) distance path between the recognised and reference transcriptions. Fig. 5.9 shows an example of minimum edit distance path between the recognised (bottom) and reference (left) transcriptions of the text line image on the bottom. As observed, each recognised word is assigned to a specific segment in the image. Then, each recognised word is assigned (if any) some edit operations, which corresponds to the supervision cases described in Section 5.4. For example, the second case corresponds to a substitution, while the third corresponds to a substitution plus one or more insertions. Therefore, levensthein operations can be employed to simulate the real user supervisions.

In the case of substitutions and deletions, these operations can be directly assigned to recognised words. For example, in Fig. 5.9, the first substitution is assigned to “sus”, the deletion assigned to “una”, and the second substitution corresponds to “camarera”. However, insertions have no direct assignment to recognised words. In our case, inserted words are assigned to the recognised word whose Viterbi segment covers most part of the Viterbi segment of the reference transcription. For instance, in Fig. 5.9, the period is completely covered by “camarera”, and thus its insertion is assumed to be done when “camarera” is supervised.

5.6.2 Interactive Experiments

In this section, we study the interactive transcription of GERMANA and RODRIGO. In the experiments, a simulated user interactively transcribes the whole document considering that the amount of effort is limited. At the end of the process, the quality of the resulting transcriptions is evaluated based on WER.

Two alternative interaction protocols have been evaluated. In both protocols, words are supervised sorted by confidence from lowest to highest. The difference is that in the first

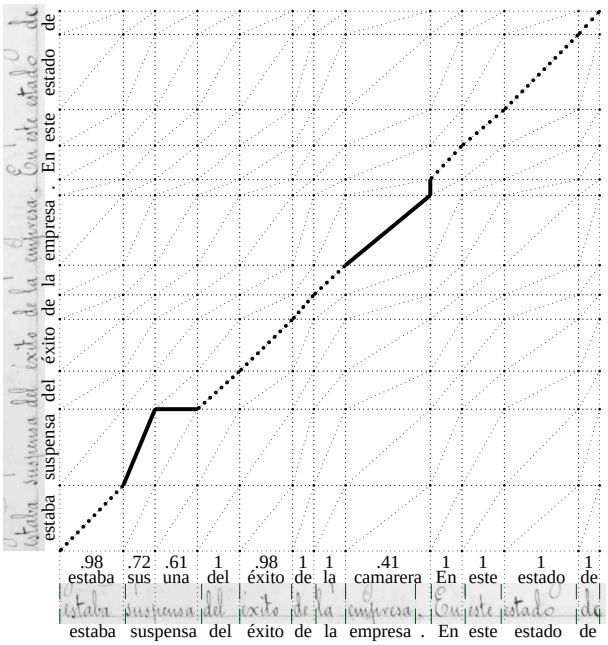


Figure 5.9: Example of minimum edit distance path between the recognised (bottom) and reference (left) transcriptions of a text line image. From bottom-left to top-right, the edit operations are, first a substitution of "sus" by "suspensa" followed by a deletion of the word "una", then, a substitution of "camarera" by "empresa" and finally the insertion of ".". On the bottom, segments of text line image are assigned to recognised words using the Viterbi alignment.

interaction protocol supervision is carried out line-per-line whereas in the second protocol supervision is performed at block level. Thus, for a given supervision effort of $X\%$, the difference is to supervise $X\%$ of the least confidence words in each line or $X\%$ of the least confident words of the block. In the first case errors are assumed to be distributed uniformly per line. Obviously, this is an unrealistic assumption but this protocol is considered since it would correspond to the usual way a document is transcribed by expert paleographers.

All the interactive learning strategies described in Section 5.4.1: conventional (C), in which no hypothesis recomputation is performed; iterative (I), in which the recomputation is performed each time a user perform a supervision; and delayed (D), in which recomputation is performed once all user supervision have been performed; have been evaluated following the line-level interaction protocol. Additionally, only the delayed strategy has also been evaluated following the block-level interaction protocol. We will denote this strategy as delayed block-level (DB). It must be kept in mind that iterative strategy fits better when supervision is performed at line-level since user attention over the whole sentence is required. Moreover, once the user has finished the supervision following the block-level interaction protocol, it seems more reasonable to apply the delayed strategy instead of conventional to (hopefully) improve the resulting transcriptions. All these strategies have been compared with the non-

interactive supervision strategy called supervised (S). In this strategy, the supervision effort of $X\%$ is employed in the manual transcription of the first $X\%$ words of the document and the rest of the document is transcribed automatically using models trained from the manual transcriptions. This last strategy is considered the baseline as it is the simplest approximation to CAT in HTR, and it does not employ any of the tools presented in this chapter is employed.

In the evaluation of interactive strategies user effort is initially devoted to fully supervise the first block (the first 1000 lines). This block is used to train and tune the initial system. This validation process is the same as in Section 4.3. In the line-level experiments, user efforts of 14%, 22%, 31% and 40% have been considered. These percentages correspond with the supervision of one, two, three or four words per line, respectively. Note that, in both corpora, the average number of words per line is 11. For the sake of comparison, the same values have been used in block-level experiments. In the case of the supervised strategy, the user effort is measured stepwise as the transcription of 2000-line blocks, which represent similar user efforts to those of the interactive experiments.

For all interactive strategies, each block is automatically transcribed and partially supervised according to each strategy. Once the supervision of one block is finished, supervised and high confidence parts of the resulting transcriptions are added as new training material to built new models to recognise the next block as explained in Section 5.5.

Figure 5.10 and Fig 5.11 shows the result of the performed experiments for GERMANA and RODRIGO, respectively. The X axis measures the user effort employed, which is calculated as the percentage of reference words that have been supervised. Word supervision is considered under the cases detailed in Section 5.4, even when it corresponds to the supervision of a correct word. In the Y axis, the quality of the transcribed document is evaluated in terms of WER.

The second point of the curves, around 56% and 50% of WER for GERMANA and RODRIGO, respectively, corresponds to the first fully-annotated block (1000 lines) used to tune all necessary parameters for interactive strategies, as shown in Sec. 4.3. It must be noted that, results are different as in this case, the system is used to recognise the remainder of the document (around 19K lines), not only the validation set defined in Sec. 4.3. Even though this system was trained from little annotated data, its evaluation provides a glimpse of the task difficulty. Both corpus have a relatively big vocabulary containing a large number of singletons. Since these words appear only once in the whole document, recognition error increases due to these out-of-vocabulary (OOV) words. This effect is greater in GERMANA, where there are six different languages and multiple document layout structures, such as list, letters, and notes.

The objective of the interactive strategies is to produce the best transcriptions with the given user effort. This best case would correspond to a curve passing as close to the XY axis as possible. On the other hand, the worst case corresponds to a diagonal line connecting the top left point, which represents a void transcription, with the bottom right point, which represents the manual annotation of the whole document. In this worst case, user effort would be devoted to manually transcribe a part of the document leaving the rest untranscribed. As observed in Figure 5.10 and Figure 5.11, all the strategies achieve to reduce user effort over manual transcription, since all curves are below the worst-case diagonal. Indeed, the same transcription quality can be achieved with lesser user effort depending on which interactive strategy is employed.

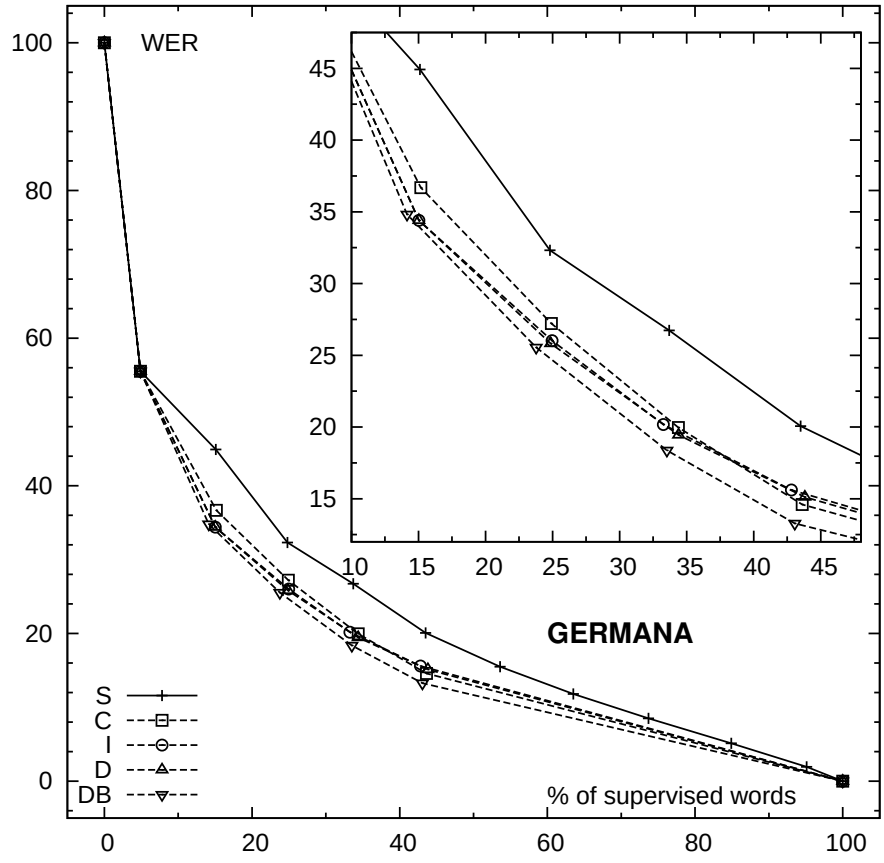


Figure 5.10: WER results from the interactive transcription experiments performed. Word Error Rate (WER) of the final transcriptions is shown for each approach using a limited user effort. A close-up is shown in the upper right corner depicting interactive approaches.

Regarding comparison between the strategies proposed, all of them present a similar behaviour. Transcription accuracy is directly related to the available user effort. However, this improvement greatly decreases when 20% of the document is supervised. This effect is caused because the initial system is not be able to deal with image character variability and language complexity. Once sufficient training data is supervised, image models are well estimated since they correspond to a unique author with a uniform script. However, the language complexity remains mostly due to OOV words. This latter effect can be directly observed in the supervised approach which improves uniformly as more data is supervised. Despite the fact that correct data improves the system as is added to the training set, the improvement from correct data is limited (Hakkani-Tür et al., 2006; Serrano et al., 2009). However, this improvement is also true in the case of interactive strategies in which data is added to the training set based on confidence measures.

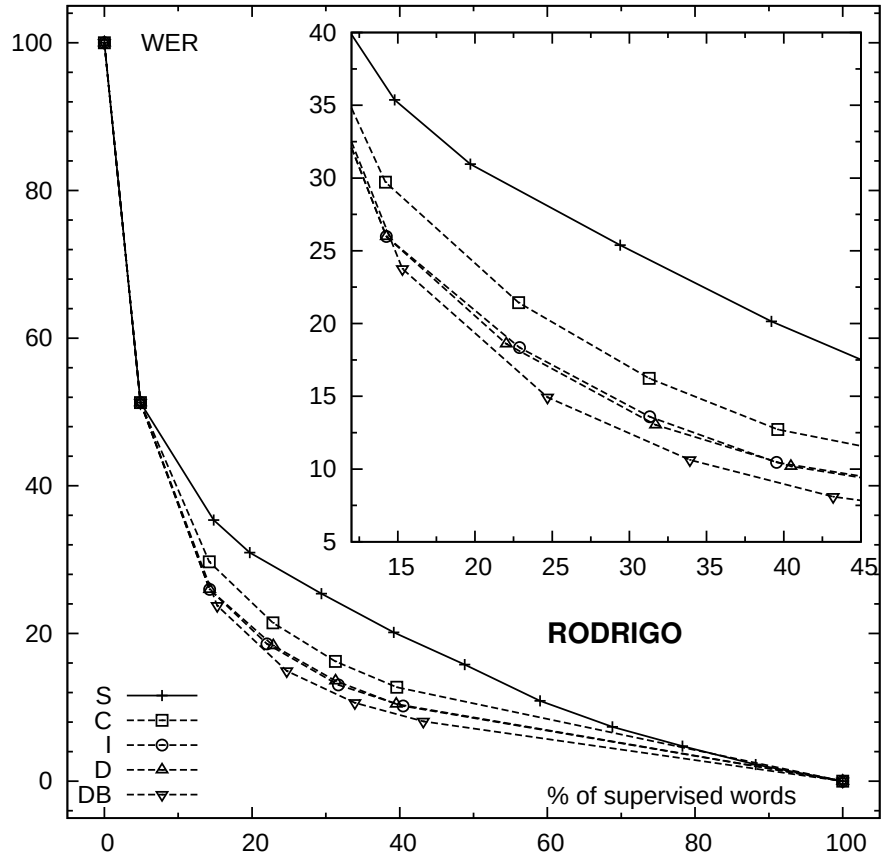


Figure 5.11: WER results from the interactive transcription experiments performed. Word Error Rate (WER) of the final transcriptions is shown for each approach using a limited user effort. A close-up is shown in the upper right corner depicting interactive approaches.

All interactive transcription strategies outperform the supervised strategy. Indeed, for a similar user effort, there is an important improvement in the transcription quality of 8, and 15 points of WER on average for GERMANA and RODRIGO, respectively. This is mainly caused because user effort is used more efficiently. Interactive strategies employ user effort to supervise likely incorrect words based on confidence measures. Consequently, user corrections directly reduces the error. On the contrary, the supervised approach supervise all words independently of their confidence which is a waste of user effort.

Performance behaviour of line level interactive approaches is slightly different from the supervised approach. There is a greater improvement in the transcription quality when the user supervises one or two words per line, with respect to the case in which three of four supervisions per line are performed. The reason behind this behaviour is an erroneous detection of incorrect words based on confidence measures, as it was shown in Figure 5.3. Confidence

measures correctly identify the first word in need of supervision 80% of the times. However the second word to be supervised is actually incorrect 60% of the times. The explanation of this difference is that, as expected, not all errors are uniformly distributed over lines. Also, small errors, such as one character mismatch, are likely to go unnoticed to the confidence measures.

Figure 5.10 and Figure 5.11 also zooms the interactive results for each corpus. In RODRIGO, both constrained search strategies, iterative (I) and delayed (D), clearly outperform the conventional (C) in all the experiments. As said, the constrained Viterbi technique, described in Section 5.4.1, recomputes the system hypothesis constrained to user supervisions. This recomputation improves the initial transcription reducing the uncertainty in the search. For example, when only one word is supervised per line, the constrained search improves the results by 5 WER points, decreasing down to 2.5 WER points when four words are supervised. This fact is directly related to the mentioned effect of the confidence measures detecting incorrect words beyond the third and fourth supervised words. On the contrary, in GERMANA, the constrained strategies only outperform the conventional strategy in 5 and 2.5 points of WER when supervising one or two words per line, respectively. A posterior analysis of the results showed that the special treatment of blank symbol described in 4.3.4 harms the constrained recomputation. As said, this treatment helps the system to recognise OOV words by the concatenation of those present in the lexicon. However, when the number of constraints is high, this feature increments the number of insertions and, thus, the number of errors.

We can also observed that there is no significant difference between the iterative and delayed strategies in both corpora when supervisions are performed on the line level, as corroborated by a bootstrap evaluation (Efron and Tibshirani, 1994). The iterative strategy was expected to be the best one since transcriptions are automatically modified based on each user supervision, resulting in a continuously guided search. However, a detailed analysis showed that the confidence of unsupervised words increase as more words are supervised and, consequently, the system recomputation does not replace them independently of their correctness. The delayed strategy can be considered as the better performance strategy because recomputation cannot be performed in real time. Long waiting times are needed in the interactive approach to recompute hypotheses. Specifically, each recomputation took 30 seconds on average in an Intel i7 with 2.80 GHz.

Regarding comparison between the two different interaction protocols, delayed block-level slightly improved all previous approaches for all user efforts considered. Concretely, results are improved by 1.25 points of WER on average. This is mainly due to a better usage of user effort which is used to supervise more erroneous words than the line-level experiments. However, the improvement is not significant in all cases and it would be expected to be higher. For instance, on the second point of GERMANA, which corresponds to a 15% of user effort on average, all approaches that include the constrained-Viterbi recomputation achieved the same result independently of the interaction protocol applied. A deep analysis of the results indicates that a uniform distribution of the error seems adequate when the available quantity of user effort is small. The reason is because, as said, the least confidence words in the lines almost correspond to the least confident words in the block. On the contrary, when supervision effort is high, uniform distribution of the error per line is unrealistic and, consequently, the block-level approach is more effective in the aim of supervising the

words which are more likely to be incorrect.

It should also be noted that there is a slight mismatch in terms of supervision effort among the interactive approaches on the line level, although the same number of supervisions are applied per line. In interactive experiments, the system may ask the user to supervise recognised words, that may not correspond to a single reference word, but two o more words. In fact, one recognised word corresponds to 1.1 reference words on average.

In the experiments discussed above user effort has been measured in terms of the percentage of supervised words. This metric has been used for two reasons. Firstly, in order to establish a fair comparison between all the strategies independently from the specific words which are supervised. Note that supervised words can be different depending on the interactive strategy applied. Secondly, the difficulty to assess user effort. Actual supervision cost can only be obtained by measuring the time cost in a real experiment with real users. This is a very cost and time consuming task and alternative metrics are needed to perform faster evaluation of the techniques. As alternative, we have considered that the percentage of supervised words is a straightforward metric which gives us an acceptable approximation to the actual cost of supervision. However, this metric has the drawback of considering the same cost for the four supervision cases detailed in Section 5.4. To circumvent this limitation, we have also used a new metric that compute the percentage of characters typed by a user in the supervision process. As a difference, this metric considers that the equal (or substitution by itself) and deletion operations have a lower edit cost than the other edit operations. Thus, equal and deletion operations only require to type one character whereas in the other supervision cases the cost is the number of characters typed by the user.

Figure 5.12 and Figure 5.13 shows the results in terms of percentage of typed characters for the baseline Supervised (S) and the best interactive approach, i.e. Delayed block-level (DB) for GERMANA and RODRIGO, respectively. As observed, the supervised approach curve in Figure 5.10 and Figure 5.12 shows the same behaviour because the user effort is employed in completely annotating the first part of the document. On the other hand, there is a great difference between the interactive approach in both figures when applying a high quantity of user effort. The interactive approach is more effective in terms of typed characters. However, the improvement achieved by using a higher user effort decreases faster than in terms of supervised words. This is mainly caused by the previously mentioned problem about the effectiveness of confidence measures. As said, the first words to be supervised are likely to be incorrect and, thus, the user has to type a higher quantity of characters. On the contrary, when more words have been supervised, supervision of correct words increases and a simple key interaction is needed for supervision. As observed in both figures, this effect greatly depended on the recognition performance. In GERMANA, as depicted in Figure 5.12, there are more errors than in RODRIGO, hence, the percentage of typed characters decreases more slowly.

In conclusion, despite the metric used to measure the supervision effort, the interactive approach outperform the baseline supervised approach for any quantity of user effort applied. However, the presented interactive approach is the most effective only when an acceptable quantity of user effort is employed. In fact, an error free transcription will require almost the same effort as the manual transcription. The effectiveness of this approach also depends on the error of the HTR system. Comparing the results from GERMANA and RODRIGO, it is shown that the interactive approach is most effective when the error is lower than 30%.

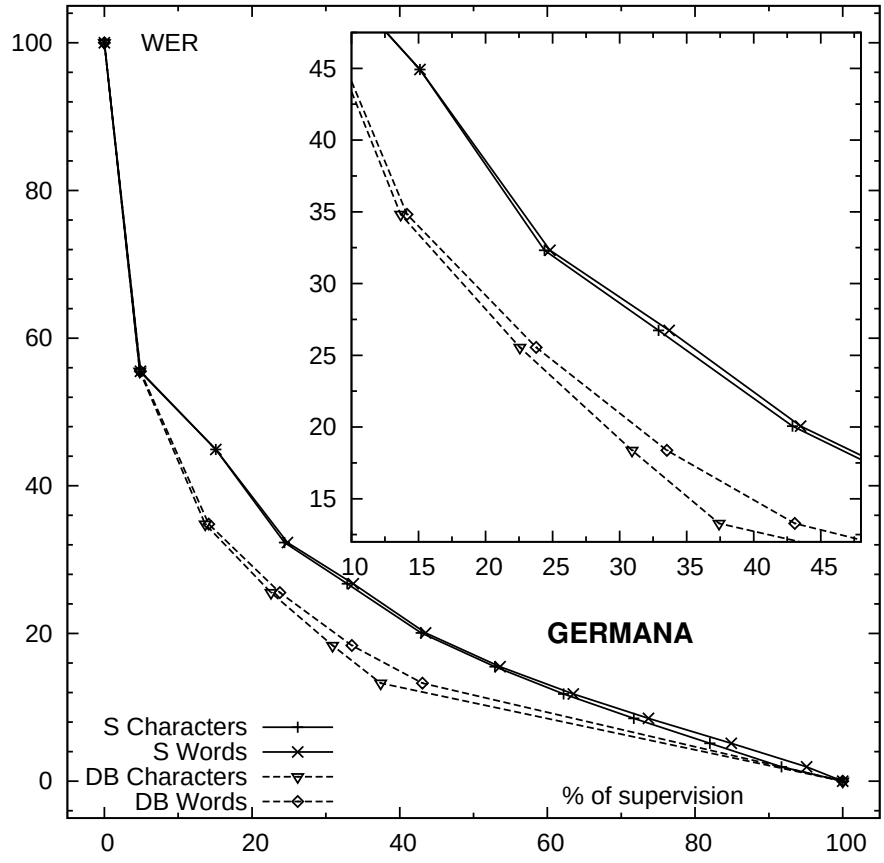


Figure 5.12: WER results from the interactive transcription experiments performed for supervised and the best interactive approaches in GERMANA. Supervision effort is measured in terms of percentage of typed characters and supervised words.

5.7 Conclusions & Future Work

In this chapter, we have described an interactive approach to handwriting text transcription when user effort is limited. This approach integrates different components that have been depicted in this chapter. First, confidences measures are used to focus user attention in those possibly incorrect words in need of supervision. Next, user supervisions are seamlessly included as constraints in the search for an alternative transcription, hopefully improving the current system hypothesis. Lastly, supervised and high confidence segments are incorporated into the training set, from which underlying image and language models are dynamically retrained. We have compared three interactive transcription strategies have been proposed to achieve an effective user interaction, that differ on how hypothesis recomputation is performed. Interactive transcription strategies have been described and their performance compared with that of a fully supervised baseline system in two real databases.

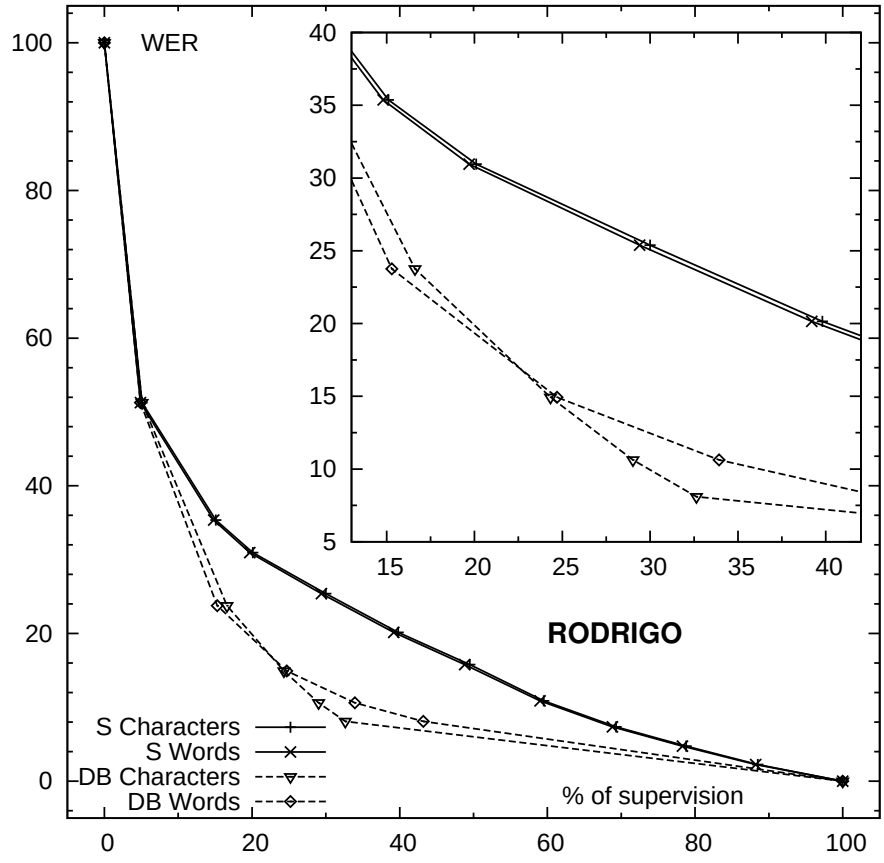


Figure 5.13: WER results from the interactive transcription experiments performed for supervised and the best interactive approaches in RODRIGO. Supervision effort is measured in terms of percentage of typed characters and supervised words.

The interactive approach proposed outperformed the baseline supervised approach for any quantity of user effort applied. However, its effectiveness strongly depends on the quantity of user effort applied. As shown, the most effective result is obtained when the user effort is low, as the detection of incorrect words decays when the least confident words have been supervised. In addition, hypothesis recomputation helped to slightly improved the transcription. Finally, the combination of active and semi-supervised learning managed to better adapt the system, and thus improve, the upcoming transcription.

User effort have been also measured in terms of the percentage of typed characters. Thus, supervision cost is different depending on the kind of supervision performed. From this point of view, interactive approaches have been more effective. However, its performance greatly decreases when supervision effort is high since the user is asked to supervise a high number of correct words. In future work, we plan to better measure the user effort using real

user evaluations on different tasks. Alternatively to a fixed number of user supervisions, in the next chapter we study the application of the interactive transcription approach presented when the user effort is variable.

The work presented in this chapter employs tools and techniques of a wide range of areas in ML, however, not all of them has been tested. For instance, some recent contributions have obtained better confidence measures from the combination of several features (Sanchis et al., 2012). This is specially appealing in our case, as from the results, the most significant results would come from a better detection of incorrect words. Alternatively, improving the hypothesis recomputation step of the process could be possible by using a different criteria. In our approach, hypothesis recomputation was performed on those words that are likely to be incorrect. However, this could not be the best criteria. For example, Culotta et al. (2006) performed hypothesis recomputation using the correction of those recognised words that would most affect its surroundings. Lastly, the system adaptation is performed using the supervised and high confident words. A minor drawback of this approach is that in the adaptation step, unsupervised recognised words can be added or not to the system, while it would be better to consider all of them in the adaptation weighted by its confidence. Using this idea, high-confident unsupervised words would contribute the most to the adaption, while low-confident would be almost ignored. Similarly, word level adaptation could be improved by adapting only those high-confident characters within the words, or continuously iterating the semisupervised adaptation process until no further improvement is detected (Wessel and Ney, 2005).

In addition, the user interaction presented in this chapter has been focused on supervision at the word level. However, user supervision at the character level may significantly reduce the effort needed to interactively transcribe a text document, specially in the presence of a large number of OOV words. For this reason, we are currently exploring this possibility to improve the performance of our interactive HTR system (Agua et al., 2012). Finally, an improved language model estimation could be obtained by successfully incorporating external resources, as explained in Section 4.3.7. For instance, selecting training samples from out-domain corpora that maximise the performance of our HTR system.

The work presented in this chapter has led to four publications in international conferences:

- L. Tarazón, D. Pérez, **N. Serrano**, V. Alabau, O. Ramos-Terrades, A. Sanchis and A. Juan. Confidence Measures for Error Correction in Interactive Transcription of Handwritten Text. *Proceedings of the 15th International Conference on Image Analysis and Processing (ICIAP 2009)* . Vietri sul Mare (Italy). Sep 2009.

- **N. Serrano**, D. Pérez, A. Sanchis and A. Juan. Adaptation from Partially Supervised Handwritten Text Transcriptions. *Proceedings of the 11th International Conference on Multimodal Interfaces and the 6th Workshop on Machine Learning for Multimodal Interaction (ICMI-MLMI 2010)*. Cambridge, MA (USA). Nov 2009.

- **N. Serrano**, A. Giménez, A. Sanchis and A. Juan. Active Learning Strategies for Handwritten Text Transcription. *Proceedings of the 12th International Conference on Multimodal Interfaces and the 7th Workshop on Machine Learning for Multimodal Interaction (ICMI-MLMI 2010)*. Beijing (China). Nov 2010.

- **N. Serrano**, A. Giménez, J. Civera, A. Sanchis and A. Juan. Interactive Handwriting Recognition with Limited User effort. *International Journal on Document Analysis and Recognition (IJDAR)*. Feb 2013.

Bibliography

M. Agua, N. Serrano, J. Civera, and A. Juan. Character-based handwritten text recognition of multilingual documents. In *Proc. of Advances in Speech and Language Technologies for Iberian Languages (IBERSPEECH 2012)*, pages 187–196, 2012.

L. V. Ahn, B. Maurer, C. Mcmillen, D. Abraham, and M. Blum. reCAPTCHA: Human-based character recognition via web security measures. *Science*, 321: 1465–1468, 2008.

S. Barrachina, O. Bender, F. Casacuberta, J. Civera, E. Cubel, S. Khadivi, A. L. Lagarda, H. Ney, J. Tomás, and E. Vidal. Statistical approaches to computer-assisted translation. *Computational Linguistics*, 35 (1):3–28, 2009.

R. Bertolami and H. Bunke. Hidden Markov model-based ensemble methods for offline handwritten text line recognition. *Pattern Recognition*, 41:3452–3460, 2008.

A. Culotta, T. Kristjansson, A. McCallum, and P. Viola. Corrective feedback and persistent learning for information extraction. *Artificial Intelligence*, 170(14):1101–1122, 2006.

B. Efron and R. J. Tibshirani. *An Introduction to Bootstrap* . Chapman & Hall/CRC, 1994.

A. Fischer, M. Wuthrich, M. Liwicki, V. Frinken, H. Bunke, G. Viehhauser, and M. Stolz. Automatic transcription of handwritten medieval documents. In *Proc. of the 15th Int. Conf. on Virtual Systems and Multimedia (VSMM 2009)*, pages 137 – 142, 2009.

D. Grangier, A. Vinciarelli, and H. Bourlard. Information retrieval on noisy text. Technical report, IDIAP, 2003.

A. Graves, M. Liwicki, S. Fernández, R. Bertolami, H. Bunke, and J. Schmidhuber. A novel connectionist system for unconstrained handwriting recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(5):855–868, 2009.

D. Hakkani-Tür, G. Riccardi, and G. Tur. An active approach to spoken language processing. *ACM Transactions on Speech and Language Processing*, 3:1–31, 2006.

T. Kristjannson, A. Culotta, P. Viola, and A. McCallum. Interactive information extraction with constrained conditional random fields. In *Proc. of the 19th National Conference on Artificial Intelligence (AAAI 2004)*, pages 412–418, 2004.

V. I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710, 1966.

L. Likforman-Sulem, A. Zahour, and B. Taconet. Text line segmentation of historical documents: a survey. *Int. Journal on Document Analysis and Recognition (IJDAR)*, 9, 2007.

E. Matusov, S. Kanthak, and H. Ney. Integrating speech recognition and machine translation: Where do we stand? In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Toulouse, France, 2006.

A. Revuelta-Martínez, L. Rodríguez, and I. García-Varea. A computer assisted speech transcription system. In *Proc. of the 13th Conf. of the European Chapter of the Association for computational Linguistics (EACL 2012)*, pages 41–45, 2012.

L. Rodríguez, I. García-Varea, and E. Vidal. Multi-modal computer assisted speech transcription. In *Proc. of the 12th Int.*

Conf. on Multimodal Interfaces and the 7th Workshop on Machine Learning for Multimodal Interaction (ICMI-MLMI 2010), pages 1–30, 2010.

R. Sánchez-Sáez, L. A. Leiva, J.-A. Sánchez, and J.-M. Benedí. Interactive predictive parsing using a web-based architecture. In *Proc. of The 11th Annual Conf. of the North American Chapter of the Association for Computational Linguistics Demonstration Session (HLT-DEMO 2010)*, pages 37–40, Stroudsburg, PA, USA, 2010.

A. Sanchis, A. Juan, and E. Vidal. A word-based naïve bayes classifier for confidence estimation in speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(2):565–574, 2012.

N. Serrano, D. Pérez, A. Sanchis, and A. Juan. Adaptation from partially supervised handwritten text transcriptions. In *Proc. of the 11th Int. Conf. on Multimodal Interfaces and the 6th Workshop on Machine Learning for Multimodal Interaction (ICMI-MLMI 2009)*, pages 289–292, 2009.

N. Serrano, A. Giménez, A. Sanchis, and A. Juan. Active learning strategies for handwritten text transcription. In *Proc. of the 12th Int. Conf. on Multimodal Interfaces and the 7th Workshop on Machine Learning for Multimodal Interaction (ICMI-MLMI 2010)*, Beijing (China), 2010.

N. Serrano, A. Giménez, J. Civera, A. Sanchis, and A. Juan. Interactive handwriting recognition with limited user effort. *Int. Journal on Document Analysis and Recognition (IJDAR)*, 2013.

B. Settles. Active learning literature survey. Technical report, 2010.

A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. Van, and M. Meteer. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26:339–373, 2000.

L. Tarazón, D. Pérez, N. Serrano, V. Alabau, O. Ramos-Terrades, A. Sanchis, and A. Juan. Confidence measures for error correction in interactive transcription of handwritten text. In *Proc. of the 15th Int. Conf. on Image Analysis and Processing (ICIAP 2009)*, pages 567–574, Vietri sul Mare (Italy), 2009.

A. Toselli, V. Romero, L. Rodríguez, and E. Vidal. Computer assisted transcription of handwritten text. In *Proc. of the 9th Int. Conf. on Document Analysis and Recognition (ICDAR 2007)*, pages 944–948, Curitiba (Brazil), 2007.

A. H. Toselli, A. Juan, D. Keysers, J. González, I. Salvador, H. Ney, E. Vidal, and F. Casacuberta. Integrated Handwriting Recognition and Interpretation using Finite-State Models. *Int. Journal of Pattern Recognition and Artificial Intelligence*, 18(4):519–539, 2004.

F. Wessel, R. Schlüter, K. Macherey, and H. Ney. Confidence measures for large vocabulary continuous speech recognition. *IEEE Transactions on Speech and Audio Processing*, 9:288–298, 2001.

F. Wessel and H. Ney. Unsupervised training of acoustic models for large vocabulary continuous speech recognition. *IEEE Transactions on Speech and Audio Processing*, 13(1):23 – 31, 2005.

X. Zhu. Semi-supervised learning literature survey, 2006.

CHAPTER 6

Balancing Error and Supervision Effort in Interactive Handwriting Recognition

Contents

6.1 Introduction

In the previous chapter, we showed how to efficiently transcribe a document when user effort is limited using an interactive approach. However, even though user effort is saved when compared to the manual transcription, it is not clear the the quality of the transcription that will be obtained when a specific user effort is applied. A more natural approach would be exactly the opposite, i.e. to apply the exact user effort needed to reach a level of quality or a user defined error rate. Ideally, in this case, the user will adjust the error desired in the final transcription, and the system will ask the user to only apply the right amount of user effort in order to obtain it. In our previous interactive system, a predefined quantity of user supervisions was applied after a text block was recognised. In the current case, the user decides which error is desired in the final transcriptions, and the system decides in the supervision degree needed to reach that error. This is easily performed if the error of a recognised block is known, as user is guided to correct the right amount of incorrectly recognised words. However, the error cannot be estimated without the reference, and it cannot be easily estimated.

The problem of estimating the error of some recognised data is typically known in the literature as accuracy or error-rate prediction. In the following, we speak in terms of Error-rate Prediction (EP), as the results on this thesis are reported in terms of error rate. EP has been typically used on practical applications. In these applications, EP estimation typically employs confidence measures to validate the system performance on a given task. For instance, Schlapbach et al. (2008a) used a EP system based on support vector regression in HTR, in which the estimation is employed to decide if a recognised text is readable enough. Similarly, Yoon et al. (2010) proposed a linear regression of multiple speech features to determine the quality of the English in real oral exams. Another application is to use the acoustic likelihood of an ASR system to better distribute the effort in a speech transcription task (Roy et al., 2010). However, these applications were not related to computer-assisted scenarios.

In this chapter, we develop a novel method to predict the error rate of automatically recognised words, and thus, estimate how much effort is required to correct a transcription to a certain user-defined error rate. The proposed method is included in the interactive approach described in the previous chapter, which efficiently employs user interactions by means of active and semi-supervised learning techniques, along with a hypothesis recomputation algorithm based on constrained Viterbi search. Transcription results, in terms of a trade-off between user effort and transcription accuracy, are reported on two real handwritten documents proving the effectiveness of the proposed approach.

The rest of this chapter is organised as follows. First, in Section 6.2, we present two error estimation algorithms depending on at which level, line or block, is the user interaction performed. Section 6.3 shows the empirical results of the proposed approach and its corresponding discussion. Finally, conclusions are drawn and future work is envisioned in Section 6.4.

6.2 Error Estimation in Automatically Recognised Words

Our objective is to estimate the WER of a set of unsupervised recognised words, whose reference transcription is unknown, in order to then decide which supervision degree is required

to reach the desired WER. Variables referring to the supervised and unsupervised parts are denoted with the plus and minus sign as superindices, respectively. Given a set of R^- unsupervised recognised words, its WER$^-$ is calculated as

$$\text{WER}^- = \frac{E^-}{N^-} \tag{6.1}$$

where E^- and N^- denote the number of editions and reference words in the unsupervised part, respectively. These variables require the reference to be know and thus cannot be used in the estimation. Assuming that errors in the supervised part occur with the same frequency as in the unsupervised part, and the ratio between recognised and reference words is also the same,

$$\frac{E^+}{R^+} \approx \frac{E^-}{R^-} \qquad \frac{R^+}{N^+} \approx \frac{R^-}{N^-} \tag{6.2}$$

Therefore, substituting our assumptions expressed in Eq. 6.2 into Eq. 6.1, we can estimate WER in the unsupervised part as

$$\text{WER}^- \approx \frac{R^- \frac{E^+}{R^+}}{R^- \frac{R^+}{N^+}} \tag{6.3}$$

In the following we present two different methods for EP in HTR, that differ on when the error is estimated. First, Section 6.2.1 describes a method that calculates the error at line level. This method was developed to be used in a line-based CAT approach, in which lines are supervised one at a time. Last, in Section 6.2.2, we present a method for EP that predicts the error on a whole block, and thus, it is intended for block-based CAT approach, in which supervisions are planned on the whole block at the beginning of the process. It must be noted that, the two approaches corresponds to the ones that were studied in the experiments of the previous chapter.

6.2.1 Line-based Prediction

Typically, manual transcription is performed line by line in the reading order. In the previous chapter, we introduced a line-based approach that was motivated by this fact, and also to avoid the user to lose the attention from a change on the context. In this approach, in order to guarantee that the error does not surpass the user defined threshold WER^*. The system, line by line, and for each recognised word in confidence order, computes the error according to Eq. 6.3. Basically, it increments in one the number of unsupervised words R^-, and the system asks the user to supervise a word when it leads to a WER^- estimate greater that WER^*.

Note that the above estimate for WER^- is pessimistic, since it assumes that, on average, correction of all unsupervised parts requires similar editing effort to that required for supervised parts, i.e. $\frac{E^+}{R^+}$. However, the user is asked to supervise recognised words in increasing order of confidence, and hence unsupervised parts should require less correction effort. In order to better estimate WER^-, we assume that errors are distributed equally across all lines, so we may group recognised words by their level of confidence c, from 1 to a certain

maximum level C, and compute a c-dependent estimate for E as above,

$$\hat{E}_c^- = \frac{E_c^+}{R_c^+} R_c^-$$

where E_c^+, R_c^+ and R_c^- are c-dependent versions of E^+, R^+ and R^-, respectively. For example, when considering four levels of confidence; $C = 1$ represents the least confident word of each line, $C = 2$ the second least, $C = 3$ the third, and $C = 4$ the rest.

The global estimate for E is obtained by simply summing these c-dependent estimates,

$$\hat{E}^- = \sum_{c=1}^{C} \hat{E}_c^-$$

and, therefore, the estimate for WER^- becomes

$$\widehat{WER}^- = \frac{\sum_{c=1}^{C} \frac{E_c^+}{R_c^+} R_c^-}{N^+ + \frac{N^+}{R^+} R^-}$$

which reduces to the previous, pessimistic estimate when only a single confidence level is considered ($C = 1$).

6.2.2 Block-based Prediction

In the experiments of the previous chapter, we concluded that a block-level approach achieved better results than its corresponding line-level counterpart. The main reason behind these results was that errors were not distributed equally across all lines. Thus, in the block-level approach, error estimation is calculated on a whole block. To better illustrate this effect, we analysed the recognition errors on the validation set of RODRIGO, as obtained in Section 4.3, to study the correctness of a recognised word depending on its confidence level. Figure 6.1 shows recognised words sorted according to their confidence measure from left (low) to right (high) in the x axis. As said, confidence measures are expected to be correlated with the correctness of each word. In this way, low confidence words are likely to be incorrect, while high confidence words are supposed to be correct. Provided the reference transcription, we are able to identify which words were incorrectly recognised, and compute the percentage of accumulated errors (y axis) in a set of words of increasing confidence. This set of words is characterised by its size, in terms of percentage with respect to the total number of recognised words (bottom x axis), or by the highest value of confidence measure in that set (top x axis). It must be noted that, these curves can be used as error estimators, as they express the error of a certain confidence interval of recognised word. Four curves representing alternative error estimators appears in Figure 6.1.

The curve labelled as *Real* assumes that the reference transcription is known beforehand, so it accounts for the accumulative percentage of errors in a set of words ordered by confidence measure. As expected, errors are more likely to occur on low confidence words, which accumulates most errors. The curve labelled as *Mean* has no access to the reference transcription and assumes that errors are uniformly distributed among recognised words, so estimating accumulative error according to Eq. 6.3. As observed, this is not an accurate error estimation.

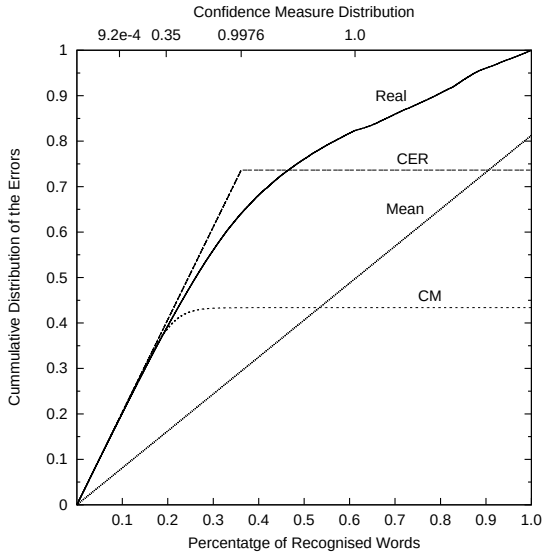


Figure 6.1: Cumulative distribution of errors on a set of recognised words ordered by confidence measure. Actual error distribution represented by the curve labelled as *Real* is compared with other error estimators based on confidence measures.

At this point, it is straightforward to consider confidence measures in error estimation. As said, confidence measures are calculated as posterior probabilities, which measure the probability of a recognised word given its corresponding word image. Similarly, one minus the posterior probability directly accounts for the probability of error of a recognised word, and it could be use as an error estimator. For instance, a recognised word with a posterior of 0.2 accounts for 0.8 errors. The curve labelled as *CM* in Figure 6.1 shows the error estimation based on the confidence measure of each word. As shown, this error estimator performs poorly if directly applied, because a large percentage of incorrect words are assigned high confidence values. Indeed, over 40% of recognised words are assigned a confidence value of one.

Alternatively, we could also consider error estimation as a classification problem, in which confidence measures are used to classify a recognised word as correct or incorrect (Schlapbach et al., 2008b). Classification is then performed by defining a threshold for confidence measures. All words below the threshold are considered incorrect, while those above are considered correct. The curve labelled as *CER* shows error estimation using a classifier based on confidence measures which threshold was adjusted to optimise the Classification Error Rate (CER) on a validation set. As shown, it also results in a poor estimation because almost 25% of errors occur over the optimised threshold, over which errors are not considered. This empirical study reveals that confidence measures cannot be directly used to predict error on a set of recognised words.

To overcome the problems previously described we proposed a novel error estimation method. This method predicts the error rate in a block of lines by estimating the number of

edit operations for each recognised word (Navarro-Cerdan et al., 2010). Given a block of R^- recognised words, let α be the ratio between the number of edit operations E^- and the number of incorrectly recognised words I^-. The α variable is motivated by the fact that an erroneous word might cause more than one edit operation, as insertions of multiple words may occur. Then, we can calculate the number of edit operations of E^- in Eq. 6.1 as

$$E^- = \alpha \mathbb{E}[I^-] \tag{6.4}$$

where $\mathbb{E}[I^-]$ is the expected value of incorrectly recognised words, since the reference transcription is not available.

Given a block of R^- recognised words, let $y_i \in \{0, 1\}$ be a random variable, which indicates if the word i is correct ($y_i = 0$) or incorrect ($y_i = 1$). Similarly, let $x_i \in \mathbb{R}$ be the confidence measure of the i-th recognised word. We assume that y_i follows a Bernoulli distribution with probability $p(y_i \mid x_i)$, i.e $y_i \sim \text{Be}(p(y_i \mid x_i))$. The number of errors I^- in a block can be estimated as

$$I^- = y_1 + y_2 + \cdots + y_{R^-} \tag{6.5}$$

and its expected value is

$$\mathbb{E}[I^-] = \mathbb{E}[y_1] + \mathbb{E}[y_2] + \cdots + \mathbb{E}[y_{R^-}] \tag{6.6}$$

Then, the expected number of errors can be calculated as

$$\mathbb{E}[I^-] = \sum_{i=1}^{R^-} \mathbb{E}[y_i] = \sum_{i=1}^{R^-} p(y_i = 1 \mid x_i) \tag{6.7}$$

Under these assumptions, the estimated number of errors in a block of recognised words is calculated as the sum of the probabilities of each word to be incorrect given its confidence measure multiplied by α. Finally, putting Eqs. 6.1, 6.2, 6.4 and 6.7 together, the estimation of WER is

$$WER^- = \frac{\alpha \sum_{i=1}^{R^-} p(y_i = 1 \mid x_i)}{R^- \frac{R^+}{N^+}} \tag{6.8}$$

Obviously, the term $p(y_i = 1 \mid x_i)$ needs to be estimated in previous blocks that have been supervised. This term can be simply calculated as

$$p(y = 1 \mid x) = \frac{N(y = 1, x)}{N(x)} \tag{6.9}$$

which is the frequency of words with confidence measure x to be incorrect.

However, the distribution of events $\{y, x\}$ is very sparse and we cannot estimate this posterior for all possible values of x. In this work, we have estimated $p(y_i = 1 \mid x_i)$ as a probability histogram, in which the domain of x is divided into a finite number of intervals.

In order to analyse the effect of the number of intervals in the accuracy of the error estimation, we performed the same experiment described in Figure 6.1 exploring the number of intervals for 1,2,8 and 32 intervals of equal size. Figure 6.2 presents a comparison of error estimation between block-based methods and the Real distribution. As observed, considering only one interval is equivalent to the mean error estimation in Eq. 6.3. Differently, each

increment of the number or intervals results in a better estimation of the error. As observed, considering 32 confidence intervals in the posterior calculation produces an accurate estimation of the error on the whole distribution. In practice, the number of intervals are optimised on a development set.

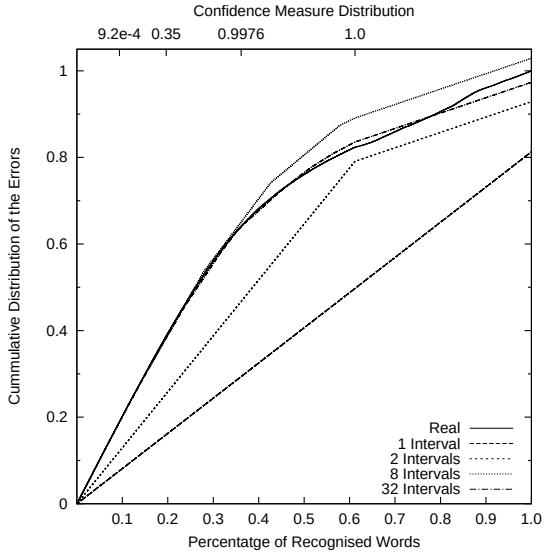


Figure 6.2: Cumulative distribution of errors on a set of recognised words ordered by confidence measure. Actual error distribution is compared with the block-based estimation studying the effect of the number of intervals.

6.3 Experiments

We performed the interactive transcription in GERMANA and RODRIGO, and compare it to a baseline, non-interactive approach. The baseline non-interactive approach (S) corresponds to an application, in which a fixed quantity of user effort is used to fully transcribe the first part of a document. Then, a HTR system is trained on this first supervised part. Finally, the rest of the document is automatically transcribed with the trained HTR system. This approach is considered to be the baseline, because it is typically the first approach applied to these tasks and no form of interactive transcription is used. On the other hand, in the interactive experiments we compared two types of error estimation approaches. First, our previous line-based method for error estimation 6.2.1. Second, the newly block-based method for error estimation that has been described in Section 6.2.2. Furthermore, as hypothesis recomputation is not considered in the error estimation as its inclusion is not straightforward, we performed an experiment to study its influence in the results. Hypothesis recomputation was presented in Section 5.4.1 in which different strategies were tested. In this chapter, we employed the best performing strategy, the *Delayed* strategy. In this strategy, hypothesis

recomputation is performed after all user interactions with the same line has been performed. The combination of error prediction methods and hypothesis recomputation results in four different approaches: line-based (L), line-based with hypothesis recomputation (L+D), block-based (B) and block-based with hypothesis recomputation (B+D).

These four approaches were employed to interactively transcribe the document given several user-defined WER thresholds for which the system balance the supervision effort required. WER thresholds were selected taking into account the average number of words per line in both documents. GERMANA and RODRIGO lines have eleven words on average due to the fact that, they have been written by a single author in well-defined templates. Then, we consider the interactive transcription of both documents when the user selects four different WER thresholds: 9% (one incorrect word per line on average), 18% (two incorrect words per line on average), 27% and 36%. It must be noted that, given that user trials are expensive and our purpose is to study the system behaviour for many different parameters, the user supervision is simulated by means of the automatic process described in Section 5.6.1.

We followed the same framework as in the previous chapter. On the one hand, in the baseline approach, we split the documents into blocks of 1000 lines. The first block is devoted to train an initial system from scratch, and tune the preprocessing, training and recognition parameters. All these optimised parameters remain unchanged for the rest of experiments. Details of this process are referred into Section 4.3. For the baseline experiment, starting from block two to last. First, we trained a system from the first to the current block and use it to recognise the rest. Finally, we measured the WER of the resulting document, i.e. on both parts, the supervised and recognised part. This experiment corresponds to a baseline non-interactive approach. On the other hand, for the interactive experiments, each database was divided into 7 consecutive blocks of 3200 lines, except for the first block, which only contains 1000 lines, and the last block, which also includes the last remnant of the lines. It should be noted that the numbers of blocks is limited in our interactive experiments due to the higher computational cost compared to the baseline. The experimental setting for each database is performed as follows. Starting from block two to the last block, each new block is processed as follows.

- First, the block is automatically recognised and confidence measures are estimated.
- Second, its recognised words are supervised according to the error estimation approach:

  Line-based approaches. As said, in Section 6.2.1, for each recognised line, words are ordered by confidence. Then, from the least confident word to the highest, the system estimates the error of all unsupervised words so far considering that the current word is not going to be supervised, which will increment the previously estimated error. If the error threshold is surpassed, the word is supervised. Four confidence intervals (C in Eq. 6.2.1) were used in all experiments. Finally, each time a word is processed, the error prediction model parameters are updated.

  Block-based approaches. The system estimates expected error on the whole block using the method presented in Section6.2.2. Then, the user supervises recognised words in order of confidence measure, independently from the line order, until the error in the remaining words is below the defined threshold. It must be noted that,

due to block segmentation of the document, the block-based approaches adjust the error on the whole document by adjusting the error independently for each block. For instance, the 9% WER threshold is achieved by adjusting the WER of all blocks to 9%.

- Third, in the approaches using hypothesis recomputation, once the user supervision is performed, the system recomputes its best hypothesis constrained to the newly supervised words and confidence measures are calculated again.

- Finally, once the whole block has been processed, it is added to the training set and the system is fully re-trained from the supervised and high-confidence words. At this step, the error prediction model of the block-based approach is also trained.

Figures 6.3 and 6.4 show the results of experiments for GERMANA and RODRIGO. On one hand, the X axis measures the quantity of supervision effort employed, which is calculated as the percentage of reference words of the document that were supervised. A word is considered to be supervised once the user is required to check that word. In fact, all four case of Section 5.6.1 count as a supervision. Note that, this includes the case of the supervision of correctly recognised words. On the other hand, the Y axis measures the quality of the produced transcriptions in terms of WER. The imaginary diagonal of these plots would represent the manual transcription of the documents. For instance, the point at coordinates $(50, 50)$ would be the result of transcribing only 50% of the document words, which will leave the rest untranscribed and it will result in 50% of WER. Similarly, the best results will correspond to a curve close to both axis, in which with the minimum effort we obtain the best transcriptions.

Each curve represents the results for each of the described interactive approaches and each point of each curve represents the result of a whole experiment. For instance, the first the line-based approach with no hypothesis recomputation in the zoomed zone of RODRIGO figure corresponds to the experiment using a user-defined WER threshold of 36%. However, due the pessimistic WER prediction described in Section 6.2.1, the resulting WER is 27%, far below the user-defined WER threshold, and the supervision effort is 21%.

As observed, all interactive approaches obtained better results than the supervised approach. It must be noted that, differences between the supervised and interactive approaches are statistically significant as shown by a bootstrap evaluation (Efron and Tibshirani, 1994). This difference is mainly caused by the combination of active and semi-supervised learning, which selects intelligently the words that have to be supervised, and then included as training data. In fact, all interactive experiments select words according to their confidence measure, which is directly related to system uncertainty. We can also observed that, as typically happens in active learning applications (Serrano et al., 2010), the improvement caused by active learning techniques decreases as the amount of user supervision available increases.

Even though all interactive approaches efficiently employ the user effort available, there are significant differences among them. The main reason behind this difference is explained by the error prediction method. As observed in both corpora, there is little difference between the supervised and the line-based approach. This is due to two problems, the ill-defined confidence intervals mentioned in Section 6.2.1, and the constraint of supervising words within a line.

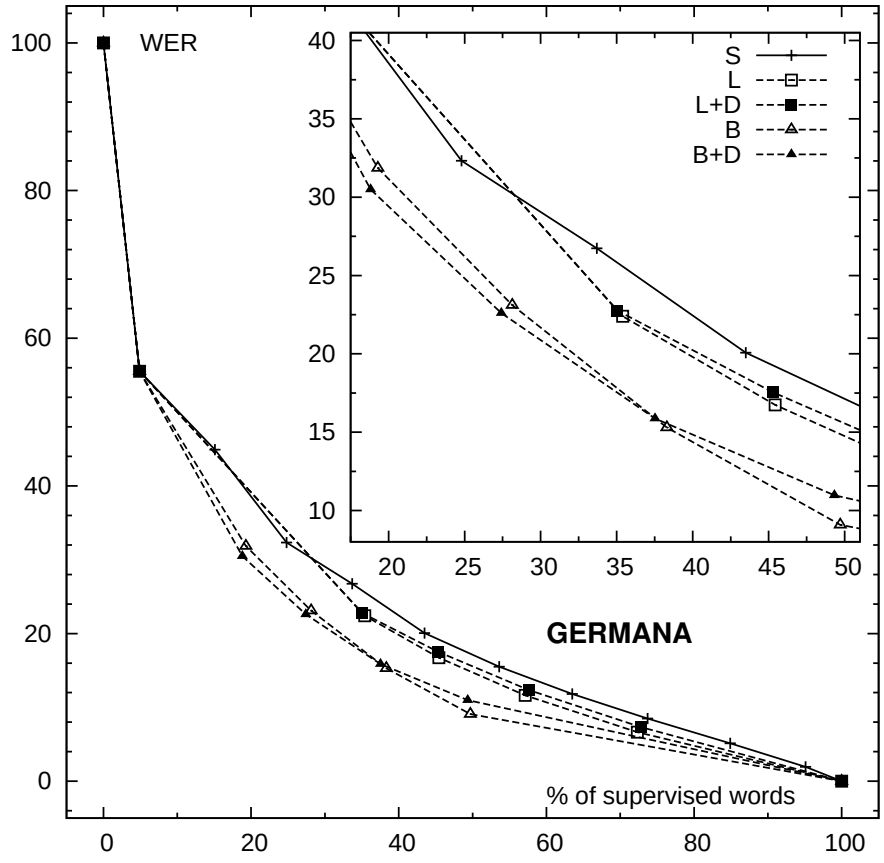


Figure 6.3: WER results from the interactive transcription experiments performed on the GERMANA database. Word Error Rate (WER) of the final transcriptions is shown for each approach using a limited user effort. A close-up is shown in the upper right corner depicting interactive approaches.

The problems of line-based approaches were overcame by two features of the newly proposed block-based approach. First, the error estimation was significantly improved by the new estimation method. Second, word supervisions are decided at block level and not constrained to line level, so better decisions can be taken to select those low confidence words inside a block.

In our experiments, as observed in Figures 6.3 and 6.4 , the block-based approach improves the line-based approach in both, system performance and efficient use of supervision effort. For instance, when comparing the supervision effort of both approaches in RODRIGO for the same transcription error. We observed that the block-based approach experiment for a WER threshold of 9% resulted in a transcription with about 9% of WER and it required a supervision effort of 51.1%. On the contrary, using the same threshold in the line-based experiment results in 7% WER and it requires a much greater amount of supervision effort,

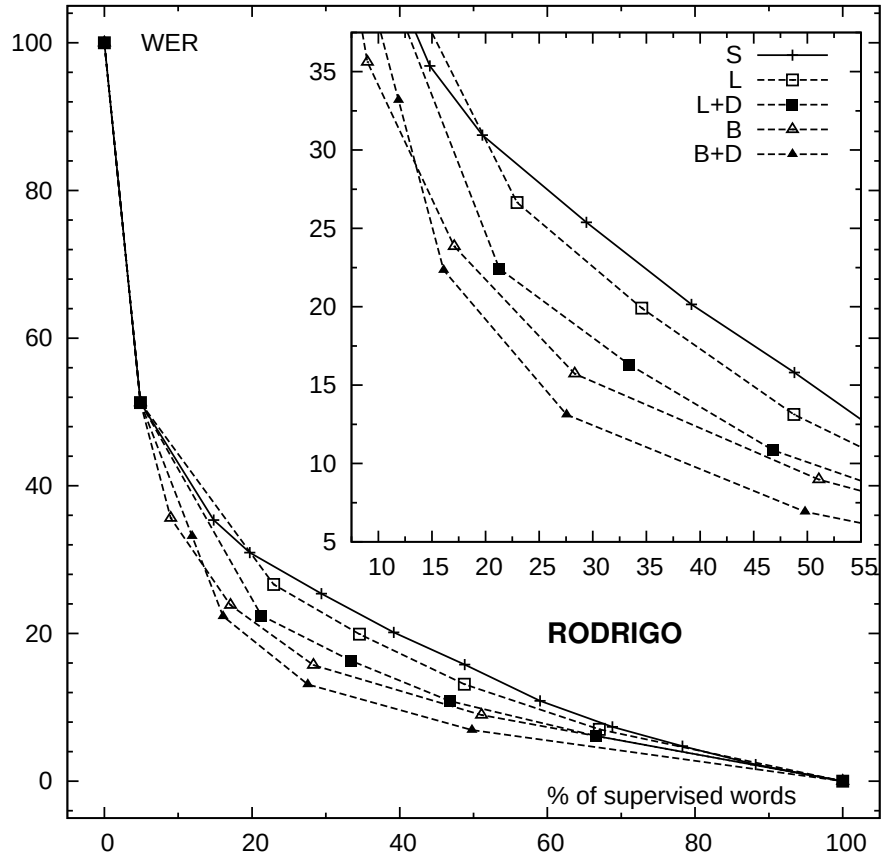


Figure 6.4: WER results from the interactive transcription experiments performed on the RODRIGO database. Word Error Rate (WER) of the final transcriptions is shown for each approach using a limited user effort. A close-up is shown in the upper right corner depicting interactive approaches.

67%. On the other hand, when comparing the error accounted by both approaches for the same supervision effort, we observed that for a supervision effort of 22.5%, the line-based approach would obtain a transcription with 27% of WER, while the block-based approach transcriptions would only contain 20% of WER. Similar improvements can also be observed in the experiments performed in GERMANA. Again, a boostrap evaluation shown that differences between the line-based and block-based results are statistically significant.

Figures 6.3 and 6.4 also include the results of both approaches when hypothesis recomputation is applied. In RODRIGO, we observe that the recomputation improves the results for both approaches in all the experiments performed. However, the improvement from this technique is much higher in the line-based approach, as the error in this approach is higher than the error of the block-based approach. In contrast, in GERMANA, it can be observed that hypothesis recomputation only improved the results slightly when supervision effort is

lower, while it performed worse when supervision effort is higher. The main cause of this behaviour is the explicit blank modelling used in GERMANA to tackle the problem of out-of-vocabulary words (OOVs). In GERMANA, as introduced in Section 4.3.4, a word is considered each time the blank character is recognised. This method is able to generate some OOVs by concatenating short words in the lexicon. However, in this case, as user supervised words are long words, constrained recognition performs words as the recogniser is tuned for obtaining short words, while the constrains correspond to much longer words. An additional problem of the hypothesis recomputation technique is that, it is not considered in the error prediction of any error estimation method. As a result, the error on final transcriptions was below the user-defined WER threshold and thus, less supervision effort could have been employed.

An additional experiment was carried out to evaluate the effectiveness of the user supervision in the best performing approach, i.e. the (B+D) approach (see Figures 6.36.4). In this experiment, we performed the interactive transcription of both documents, but considering the case in which the user adjusted the amount of user effort available instead of the WER threshold.

In this scenario, the objective of the system is to generate the best possible transcriptions with the amount of user effort available. Here we followed the same interactive approach except for the error estimation method. Instead, the decision of which words were supervised was taken by uniformly distributing the user effort available across blocks. Then, for each block, the system asked the user to supervise the corresponding least confident words. Hence, the results obtained with this approach can be directly compared with those obtained, as the only difference is the user effort applied on each block.

It should be noticed that the approach presented so far in this thesis applies a variable number of supervisions per block depending on the estimated error within the block. However, the latter approach uniformly distributes the user effort available among all blocks. As a result, a comparison between a fixed and a variable number of supervisions can be performed. The results of transcribing both corpora, GERMANA and RODRIGO, using the best approach (B+D) with the same error threshold, and using the previously presented fixed user effort approach (U), when supervising the first block and a $\{10\%, 20\%, 30\%, 40\%\}$ of the remainder blocks, is depicted in Figure 6.5.

As observed, the curves of both approaches overlap, from which we can draw two conclusions. First, the interactive transcription approach is effective for cases in which either the error or the user effort is fixed. Secondly, even though a fixed and a variable number of supervisions per block achieved similar results in terms of WER and percentage of supervised words, there are notable differences in the number of incorrectly supervised words. A further analysis revealed that, the based on, i.e. a variable number of supervisions, supervises more incorrect words than the uniform approach, as the supervision degree is higher for the first blocks when the system is still learning. On the contrary, in the case of a fixed number of supervisions per block, when the last blocks are processed and the system is better trained, the system is more likely to ask the user to supervise correct words, which wastes the available user effort.

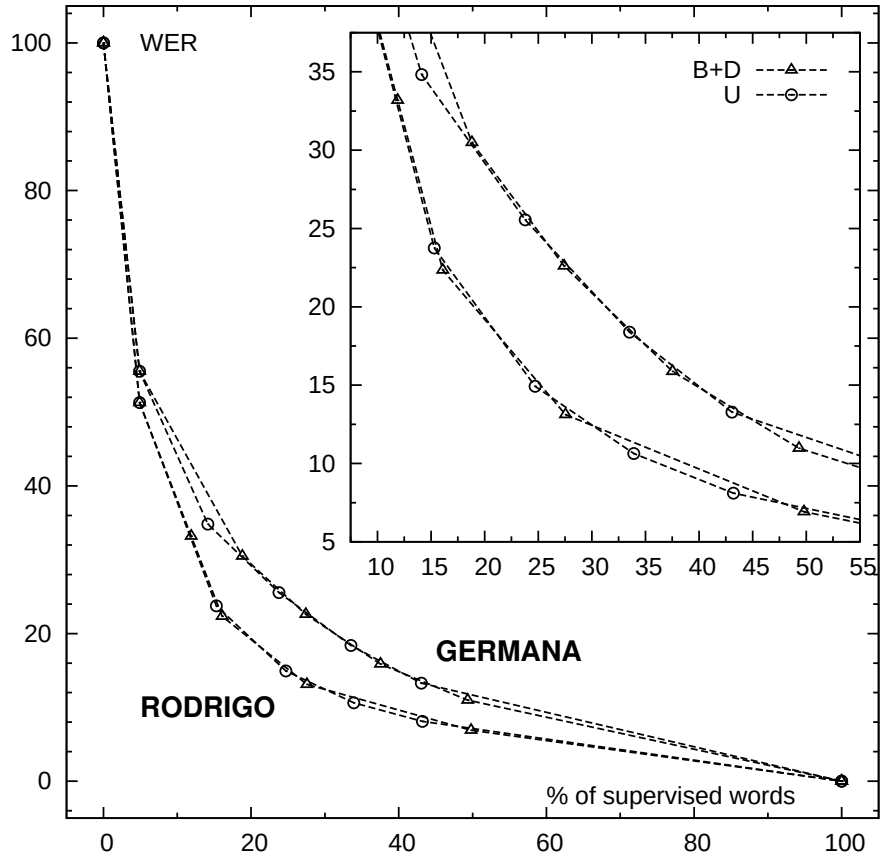


Figure 6.5: WER results from the interactive transcription experiments performed on the GERMANA and RODRIGO databases. Word Error Rate (WER) of the final transcriptions is shown for each approach using a limited user effort. A close-up is shown in the upper right corner depicting the results.

6.4 Conclusions & Future Work

In this chapter, we have presented a CAT approach to HTR when a user-defined amount of error is adjusted. We proposed two methods to estimate the WER of a set of recognised words. These methods estimate the expected number of edit operations of a recognised word by calculating the expected error of a word subjected to its confidence measure. The first method was developed to be used on a line-based approach, while the second operates at the block level. The error estimation method is included in a CAT approach that efficiently employs a limited amount of user effort by means of active and semi-supervise learning techniques, along with hypothesis recomputation to include user supervision as new search constraints.

Experiments were performed in the transcription of two real handwritten text documents.

The results obtained confirm the correctness of this approach, as the error of the transcriptions produced is always under the user defined threshold. However, the block-based approach significantly superseded the line-based approach in both, system performance and user effort reduction. In fact, the error estimation obtained with the block-level approach is close to the user defined, and it is achieved with the minimum amount of user effort that is possible using this CAT framework.

We also measured the improvement due to hypothesis recomputation when user supervisions are performed. Hypothesis recomputation improved WER results, however as words that will be corrected due to hypothesis recomputation are not considered in our error estimation method, they employed more user effort that would be required. Taking into consideration the contribution of hypothesis recomputation in the error estimation method could be achieved by using information theory metrics as was shown by Culotta et al. (2006).

On the other hand, even though an accurate error estimation was performed on the block-based approach, further analysis revealed that the proposed method may be pessimistic because of the training data used. This is caused by the fact that, training data are biased because most of supervised words are low confident words, so the error on high confident words is not being re-estimated. A better idea would be to make a better selection of the training data to estimate an error distribution similar to that of the next block. Similarly, a uniform supervision of the newly recognised words could be employed to adapt the error estimation parameter, as a linear transformation from the current error estimation function. In fact, it will be sufficient to supervise a few words from all confidence intervals and used them to refine the current error estimation. However, this supervision will take user effort, thus, a trade-off between the improvement in the estimation and the increment of the user effort should be achieved. Also, an online adaptation of the error estimation parameters each time a word is supervised, as it is performed in the line-based approach, could be useful in some applications and also remains as future work.

The work presented in this chapter has led to a publication in an international conference and a publication in an international journal:

- **N. Serrano**, A. Sanchis and A. Juan. Balancing Error and Supervision Effort in Interactive-Predictive Handwriting Recognition. *In Proceedings of the 15th International Conference on Intelligent User Interfaces*. Hong Kong (China). Feb 2010.

- **N. Serrano**, J. Civera, A. Sanchis and A. Juan. Effective balancing error and user effort in interactive handwriting recognition. *Pattern Recognition Letters*. March 2013.

Bibliography

A. Culotta, T. Kristjansson, A. McCallum, and P. Viola. Corrective feedback and persistent learning for information extraction. *Artificial Intelligence*, 170(14):1101–1122, 2006.

B. Efron and R. J. Tibshirani. *An Introduction to Bootstrap* . Chapman & Hall/CRC, 1994.

J. R. Navarro-Cerdan, J. Arlandis, J.-C. Perez-Cortes, and R. Llobet. User-defined expected error rate in OCR postprocessing by means of automatic threshold estimation. In *Proc. of the 2010 12th Int. Conf. on Frontiers in Handwriting Recognition (ICFHR 2010)*, pages 405–409, 2010.

B. Roy, S. Vosoughi, and D. Roy. Automatic estimation of transcription accuracy and difficulty. In *Proc. of the 11th Annual Conf. of the Int. Speech Communication Association (INTERSPEECH 2010)*, pages 1902–1905, 2010.

A. Schlapbach, F. Wettstein, and H. Bunke. Estimating the readability of handwritten text - a support vector regression based approach. In *Proc. of the 20th Int. Conf. on Pattern Recognition (ICPR 2008)*, pages 1–4, 2008a.

A. Schlapbach, F. Wettstein, and H. Bunke. Automatic estimation of the readability of handwritten text. In *Proc. of the 16th European Conf. on Signal Processing (EUSIPCO 2008)*, pages 2–6, 2008b.

N. Serrano, A. Giménez, A. Sanchis, and A. Juan. Active learning strategies for handwritten text transcription. In *Proc. of the 12th Int. Conf. on Multimodal Interfaces and the 7th Workshop on Machine Learning for Multimodal Interaction (ICMI-MLMI 2010)*, Beijing (China), 2010.

S.-Y. Yoon, L. Chen, and K. Zechner. Predicting word accuracy for the automatic speech recognition of non-native speech. In *Proc. of the 11th Annual Conf. of the Int. Speech Communication Association (INTERSPEECH 2010)*, pages 773–776, 2010.

CHAPTER 7

Conclusions

Contents

7.1 Summary

The work developed in this thesis has covered the whole process of interactively transcribing a handwritten text document. In Chapter 3, we introduced the interactive annotation process from a theoretical point of view. The interactive process was divided into different steps corresponding to assumptions on the optimum solution to the task. From the general interactive annotation process two different applications were presented. One focused on the interactive transcription of old text documents, and another on the interactive document layout analysis.

Chapter 4 details the acquisition and annotation process of two old text documents, and how two freely available databases called GERMANA and RODRIGO were generated from them. These databases were built because of the lack of similar resources in order to develop the CAT approach of this thesis. GERMANA and RODRIGO were carefully selected to reflect the challenges of HTR. We also described the construction of a baseline system used in the CAT approach along with results for a baseline fully supervised approach on both corpora.

The next chapters were focused on the CAT approach developed, which is the main contribution of this thesis. In Chapter 5, we presented a new approach for CAT when user effort is limited, and hence, the complete transcription of a document is not required. This approach was developed as a combination of techniques of well defined areas from ML built on top of the developed GIDOC prototype. It consists in three steps. First, the limited user effort is used to supervise possibly incorrect words that are identified thanks to CMs. Next, a constrained Viterbi recomputation is performed to improve the current system hypothesis from the newly available user supervisions. Finally, the system is adapted from supervised and high confident words due to a combination of active and semisupervised learning techniques. The effectiveness of this approach was empirically demonstrated on the transcription of GERMANA and RODRIGO, specially when the amount of user effort available is small. Experiments were carried out by a simulated user to exhaustively test this approach. It must be noted that, to our knowledge this is the first time that all these techniques are included in HTR.

Finally, in Chapter 6, we extended the previous approach to dynamically adjusting the quantity of user effort required for the task. In this new approach, the user rather than adjusting the supervised degree, he or she adjusts the error desired in the final transcriptions. We developed two methods to estimate the error on a recognised transcription, and thus, calculate the effort required for its partial correction. Again, the developed approach was tested on the transcription of GERMANA and RODRIGO on the CAT system developed in Chapter 5. Results showed that error estimation is accurate and leads to an optimum use of the effort required.

In addition, a prototype for interactive transcription called GIDOC is presented in Appendix A. GIDOC is a first step to enable transcribers to use a CAT approach for carrying out their work. Its main contribution is that it has been designed to free non HTR experts of the details of the system implementation. GIDOC is built as a set of GIMP plug-ins that deals with different parts of the whole transcription process. For instance, it includes projection-based algorithms to detect text blocks and lines within images; and a built-in software to train a standard HTR system. Furthermore, it includes an interactive interface to recognise text line images and highlight its possibly erroneous words. Thus, relieving the user from the tedious

transcription task. It is worth noting that the software is freely available under a GPL license.

Summing up, the main contributions of this thesis are:

1. The theoretical formulation of the interactive annotation of sequential data, and its application on two different tasks: the interactive transcription and the document layout analysis of old text documents.

2. The generation of two databases for HTR called GERMANA and RODRIGO. We described the acquisition and annotation process of two old text documents. Both documents present the typical problems for HTR, i.e. a difficult language structure and high number of OOV words, and are released to the community for future comparison. We depicted the construction of a baseline system and extracted empirical results.

3. A CAT approach for efficient transcription with limited user effort. This approach efficiently employs user supervision by first, asking the user to correct possibly incorrect words. Incorrectly recognised words are identified by means of CMs extracted from recognised words. Next, the supervised transcription is further improved by means of a constrained-Viterbi hypothesis recomputation to user supervisions. Finally, system models are adapted from supervised and high confidence unsupervised words. The effectiveness of this approach is empirically showed in the transcription of GERMANA and RODRIGO.

4. An approach to balance the recognition error and supervision effort. Methods to estimate the error on a set of recognised words have been developed and presented. These methods are used to estimate the supervision degree needed to achieve a transcription with a user adjusted error. Experiments showed the correctness of this approach in terms of error estimation accuracy and transcription results.

5. A prototype for interactive transcription called GIDOC. It is implemented as a set of GIMP plug-ins. GIDOC includes tools and techniques for building state-of-the-art HTR systems, and it is oriented to non-experts users, freeing them from technical details. The prototype is freely available under GPL license.

7.2 Scientific Publications

Several articles have been written in the development of this thesis, and they have been published in international workshops and conference. In this section, we briefly review these publications and their relation with the work developed in this thesis.

The interactive pattern recognition theory presented in Chapter 3 was applied to DLA and produced a publication in an international conference:

- O. Ramos, **N. Serrano** and A. Juan. Interactive-predictive detection of handwritten text blocks. In *Proceedings of the 17th Document Recognition and Retrieval Conference (DRR 2010)*. San Jose (USA). January 2010.

The two databases that has been described in Chapter 4 and have been used in the experiments in this thesis, have been published in two international conferences:

- D. Pérez, L. Tarazón, **N. Serrano**, F. Castro, O. Ramos and A. Juan. The GERMANA database. In *Proceedings of the 10th International Conference on Document Analysis and Recognition (ICDAR 2009)*. Barcelona (Spain). July 2009.
- **N. Serrano**, F. Castro and A. Juan. The RODRIGO database. In *Proceedings of 7th Language Resources and Evaluation Conference (LREC 2010)*. Valletta (Malta). May 2010.

The interactive transcription approach presented in Chapter 3 has led to three publications in international conferences and a publication in an international journal:

- L. Tarazón, D. Pérez, **N. Serrano**, V. Alabau, O. Ramos-Terrades, A. Sanchis and A. Juan. Confidence Measures for Error Correction in Interactive Transcription of Handwritten Text. In *Proceedings of the 15th International Conference on Image Analysis and Processing (ICIAP 2009)* . Vietri sul Mare (Italy). September 2009.
- **N. Serrano**, D. Pérez, A. Sanchis and A. Juan. Adaptation from Partially Supervised Handwritten Text Transcriptions. In *Proceedings of the 11th International Conference on Multimodal Interfaces and the 6th Workshop on Machine Learning for Multimodal Interaction (ICMI-MLMI 2010)*. Cambridge, MA (USA). November 2009.
- **N. Serrano**, A. Giménez, A. Sanchis and A. Juan. Active Learning Strategies for Handwritten Text Transcription. *Proceedings of the 12th International Conference on Multimodal Interfaces and the 7th Workshop on Machine Learning for Multimodal Interaction (ICMI-MLMI 2010)*. Beijing (China). November 2010.
- **N. Serrano**, A. Giménez, J. Civera, A. Sanchis and A. Juan. Interactive Handwriting Recognition with Limited User effort. *International Journal on Document Analysis and Recognition (IJDAR)*. February 2013.

Finally, the balancing approach presented in Chapter 6 has produced a publication in an international conference, and a publication in an international journal:

- **N. Serrano**, A. Sanchis and A. Juan. Balancing Error and Supervision Effort in Interactive-Predictive Handwriting Recognition. *In Proceedings of the 15th International Conference on Intelligent User Interfaces*. Hong Kong (China). February 2010.
- **N. Serrano**, J. Civera, A. Sanchis and A. Juan. Effective balancing error and user effort in interactive handwriting recognition. *Pattern Recognition Letters*. March 2013.

The prototype presented on the Appendix A has led to a publication in an international workshop:

- **N. Serrano** and L. Tarazón and D. Pérez and O. Ramos-Terrades and A. Juan. The GIDOC prototype. In *Proceedings of the 10th International Workshop on Pattern Recognition in Information Systems (PRIS 2010)* , Funchal (Portugal) June 2010.

Also, not directly product of this thesis, derived work have been employed in other publications related to HTR and interactive transcription of documents:

- M. del Agua, **N. Serrano**, J. Civera and A. Juan. Character-Based Handwritten Text Recognition of Multilingual Documents. In *Proceedings of IBERSPEECH 2012*. Madrid (Spain). November 2012.
- A. Toselli, **N. Serrano**, A. Giménez, I. Khoury, A. Juan, E. Vidal. Language Technology for Handwritten Text Recognition. In *Proceedings of IBERSPEECH 2012*. Madrid (Spain). November 2012.
- I. Sanchez, **N. Serrano**, A. Sanchis, A. Juan. A prototype for Interactive Speech Transcription Balancing Error and Supervision Effort. In *Proceedings of the 2012 ACM international conference on Intelligent User Interfaces (IUI 2012)*. Lisbon (Portugal). February 2012.
- L. Leiva, V. Alabau, V. Romero, F. Segarra, R. Sanchez, D. Ortiz, L. Rodríguez, **N. Serrano**. Prototypes and Demonstrators. Chapter of the book *Multimodal Interactive Pattern Recognition and Applications*. Springer. 2012.
- **N. Serrano**, A. Giménez, A. Sanchís and A. Juan. Active Interaction and Learning in Handwritten Text Transcription. Chapter of the book *Multimodal Interactive Pattern Recognition and Applications*. Springer. 2012.
- V. Romero, J. Andreu, **N. Serrano**, E. Vidal. Handwritten Text Recognition for Marriage Register Books. In *Proceedings of the 11th International Conference on Document Analysis and Recognition (ICDAR 2011)*. Barcelona (Spain). September 2011.
- M. del Agua, **N. Serrano**, A. Juan. Language Identication for Interactive Handwriting Transcription of Multilingual Documents. In *Proceedings of the 5th Iberian Conference on Pattern Recognition and Image Analysis*. Palma de Gran Canaria (Spain). June 2011.
- V. Romero, **N. Serrano**, A. Hector, J. Andreu and E. Vidal. Handwritten Text Recognition for Historical Documents. In *Proceeding of the 1rst Workshop on Language Technologies for Digital Humanities and Cultural Heritage (LaTeCH 2011)*. Portland (USA). June 2011.
- A. Juan, V. Romero, J. Andreu, **N. Serrano**, A. Hector and E. Vidal. Handwritten Text Recognition for Ancient Documents. In *Proceeding of the 1rst Workshop on Applications of Pattern Analysis (WAPA 2010)*. London (United Kingdom). September 2010.

7.3 Future Work

As the work presented in this thesis covers the whole interactive transcription process of a document and it employs several techniques from sub-areas of PR. There are many research lines that would be interesting to explore as a future work.

In Chapter 3, several assumptions were performed to deal with the estimation of the interactive annotation model presented. An important assumption was the division of the

whole process in several independent steps. First, select samples to be supervised, next, update the system hypothesis constrained to the user supervision, and finally, adapt the system with the new information acquired. In future work, we plan to study how to integrate all the steps into a single one, in order to improve the whole process, as all steps will be dependent. For instance, samples selected will depend on how they will influence in the recomputation and the model adaptation.

In Chapter 4, we presented two databases for HTR along with baseline results. These results showed that there is still an important room for further improvement. A possible improvement would be to try out other approaches rather than HMMs, such as NN (Graves et al., 2009) or tandem systems (Kozielski et al., 2013). Furthermore, as seen in Section 4.3.7, we have studied the impact of using an external resource, Google n-grams, in the training of the system. However, some recent contribution have managed to improve the system performance by means of adding several external resources (Valor et al., 2012) (Wuthrich et al., 2009).

CM have been extensively used in this thesis and their refinement will improve the overall performance of the interactive approaches presented. For instance, Sanchis et al. (2012) proposes the estimation of a CM using Naive Bayes classifier that combines multiple features extracted from the recognised words. They also showed that even in case of only using one feature, which coincides with the CM used in this thesis, the resulting CM improved due to a normalisation applied by the Naive Bayes classifier. Paralelly, our systems asks words in order of confidence, as proposed by the least sampling technique of AL. Nevertheless, AL is a well studied area and it could be possible to find a more suitable supervision strategy (Settles, 2010) depending on the application. For instance, instead of least sampling, recognised words could be chosen of those which most will improve in a posterior adaptation step.

Another important step on our system is the hypothesis recomputation constrained to user supervision. Our results showed that this technique slightly improved the baseline results. Specifically, the constrained user supervisions were those resulting from the CM employed, which (mostly) correspond to incorrectly recognised words. However, this latter technique is not related with the recomputation, and thus, the supervision of other words might lead to better final results, as showed by Culotta et al. (2006). For instance, in a sentence with three errors, the correction of the lowest confident word and the posterior hypothesis recomputation may correct two of this errors. However, correcting a different word, with a greater CM, may correct all errors. In fact, this behaviour was observed when comparing the iterative and delayed methods discussed in Figure 5.6. Furthermore, our current strategy selection, trying to find out the incorrect ones using CMs, and the previously proposed one, selecting those that most improve the recomputation, could be effectively merged in order to find an effective trade-off.

The last step of our approach was the system adaptation from partially supervised words. This adaptation is performed by re-training the whole model with those segments that are considered correct, specifically the user supervised ones and the unsupervised high confident ones. Nevertheless, this adaptation has three major drawbacks. First, segments are build up from words, when it could be more adequate to consider smaller segments, such as characters or even only part of them (Wessel and Ney, 2005). This is motivated by the fact that, incorrect words are similar to their reference in terms of characters, for instance “lago” and “pago”. Second, recognised words can only be considered correct or incorrect in the adaptation, when

it could be better to perform a CM weighted adaptation from them. This is similar to the EM learning algorithm (qi Han et al., 2009), in which each sample contributes the parameter estimation weighted by a factor measuring its importance. Last, adaptation was performed as a complete retraining, once a whole block of text lines was supervised. This is mainly due to the fact that the computational cost of of training a recognising. However, a better approach would be to perform an online adaptation of the system, each time a supervision is committed (Ortiz-Martínez et al., 2010), and only re-train from scratch once sufficient new data is acquired.

Our last contribution was a method to estimate the error on a set of recognised words. A problem of this approach is that it is estimated from the past model performance, thus, it is pessimistic on the recognition of new samples in which the model have been slightly improves by the adaptation. This problem could be solved by means of an information theory metric (Culotta et al., 2006), measuring how much the model have improved. Another problem of our approach is that only low confidence words are supervised, which introduces a bias on the error estimation adaptation, as only the estimation on low confidence words is updated. A possible solution to this problem is to a small uniform supervision of all supervised words, in order to adapt high confidence segments. Finally, in our error estimation, the improvement derived from the hypothesis recomputation is not considered in the error estimation, resulting in a more pessimistic estimation. This improvement could be integrated by means of the mutual information, as shown in (Culotta et al., 2006), which measures how much a recognised words depends on the other words of the sentence.

Our CAT approach has been only tested on the developed databases GERMANA and RODRIGO. In order to validate its application, it would be interesting to test it in another databases, such as IAM (Marti and Bunke, 2002) or ESPOSALLES (Romero et al., 2012). In addition, the technology used in this work derives from the technology on ASR, and thus, its application is direct. Even though, Sánchez-Cortina et al. (2012) has applied some of the techniques in this thesis, the complete approach has not been tested yet. In the moment of writing this thesis, the interactive approach proposed in this thesis is being applied to the transcription of video lectures within the transLectures project (transLectures). One of the objectives of this project is to develop cost-effective solutions to transcribe lectures recorded in universities, as annotation of these resources is plan to be perform voluntarily. The interactive approach presented on this thesis is adequate to this application, as with a low quantity of user effort, a great quality transcription could be obtained. Some preliminary results are detailed in (translectures-wp4-m12).

Much work remains to turn the GIDOC prototype presented in Appendix A into a fully operational tool. Exhaustive tests with real users have to be carried out to solve usability problems. Paralelly, a GIDOC library could be implemented, in which each functionality of GIDOC is documented and offered as a stand-alone function, to enable the community to easily extend the software. Unfortunately, GIDOC technology has became outdated. For instance, GIDOC feature extraction method was used in our baseline system until we integrated the feature extraction method of Dreuw et al. (2011), which improved this important step. Similarly, neural network based (Graves et al., 2009) (España-Boquera et al., 2011) (Hinton et al., 2012) (Kozielski et al., 2013) systems have recently outperformed Gaussian HMMs, which are GIDOC baseline, becoming the state-of-the-art. It remains as a future work to include this technology in GIDOC and evaluate it in an interactive framework.

Finally, all interactive experiments performs have been tested with a simulated user. Real user experiments remain to be done, further validating the proposed approach. Real experiments may reveal that some user effort has to be measured with a different metric, e.g. supervision of correctly recognised words does not cost the same that incorrectly recognised words. Furthermore, even though experiments with real users have sometimes shown that interactive approaches were meaningless (Luz et al., 2008), we are certain that this approach will be effective under certain assumptions. For instance, H.Nanjo and T.Kawahara (2006) show that a recognised transcription should have at most 25% of error for an interactive transcription approach to be applied. This is in case or standard interactive approach which finality is to completely supervise a transcription. In our approach, applying a lower effort, we could reach a non-perfect transcription similar to one obtained with a non-professional transcribe.

Bibliography

A. Culotta, T. Kristjansson, A. McCallum, and P. Viola. Corrective feedback and persistent learning for information extraction. *Artificial Intelligence*, 170(14):1101–1122, 2006.

P. Dreuw, G. Heigold, and H. Ney. Confidence and Margin-Based MMI/MPE Discriminative Training for Offline Handwriting Recognition. *International Journal on Document Analysis and Recognition (IJDAR)*, 14(3):273–288, 2011.

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

A. Graves, M. Liwicki, S. Fernández, R. Bertolami, H. Bunke, and J. Schmidhuber. A novel connectionist system for unconstrained handwriting recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(5):855–868, 2009.

G. Hinton, L. Deng, D. Yu, G. Dahl, A. rahman Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury. Deep neural networks for acoustic modeling in speech recognition. *Signal Processing Magazine*, 2012.

Y. H.Nanjo and T.Kawahara. Computer assisted speech transcription system for efficient speech archive. In *Proc. of the 2006 Western Pacific Acoustics Conference (WESPAC 2006)*, 2006.

M. Kozielski, P. Doetsch, and H. Ney. Improvements in rwth's system for off-line handwriting recognition. pages 935–939, 2013.

S. Luz, M. Masoodian, and B. Rogers. Interactive visualisation techniques for dynamic speech transcription, correction and training. In *Proc. of the 9th Int. Conf. on Human-Computer Interaction (CHINZ 2008)*, pages 9–16, Wellington, New Zealand, 2008.

U. V. Marti and H. Bunke. The IAM-database: an English sentence database for off-line handwriting recognition. *International Journal on Document Analysis and Recognition (IJDAR)*, pages 39–46, 2002.

D. Ortiz-Martínez, I. García-Varea, and F. Casacuberta. Online learning for interactive statistical machine translation. In *Proceedings of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL HLT)*, pages 546–554, 2010.

H. qi Han, D.-H. Zhu, and X. feng Wang. Semi-supervised text classification from unlabeled documents using class associated words. In *Proc. of the Int. Conf. on Computers Industrial Engineering (CIE 2009)*, pages 1255–1260, 2009.

V. Romero, A. Fornés, N. Serrano, J. Andreu, A. Toselli, V. Frinken, E. Vidal, and J. Llados. The esposalles database: An ancient marriage license corpus for off-line handwriting recognition. *Pattern Recognition*, 2012.

I. Sánchez-Cortina, N. Serrano, A. Sanchis, and A. Juan. A prototype for interactive speech transcription balancing error and supervision effort. In *Proc. of the 2012 ACM Int. Conf. on Intelligent User Interfaces (IUI 2012)*, pages 325–326, 2012.

A. Sanchis, A. Juan, and E. Vidal. A word-based naïve bayes classifier for confidence estimation in speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(2):565–574, 2012.

B. Settles. Active learning literature survey. Technical report, 2010.

transLectures. transLectures: Transcription and Translation of Video Lectures. `http://www.translectures.eu/`.

translectures-wp4-m12. First report on intelligent interaction of the transLectures project . `http://www.translectures.eu/progress`.

J. Valor, A. Pérez, J. Civera, and A. Juan. Integrating a state-of-the-art asr system into the opencast Matterhorn platform. In *Proc. of Advances in Speech and Language Technologies for Iberian Languages (IBERSPEECH 2012)*, pages 237–246. Madrid (Spain), nov 2012.

F. Wessel and H. Ney. Unsupervised training of acoustic models for large vocabulary continuous speech recognition. *IEEE Transactions on Speech and Audio Processing*, 13(1):23 – 31, 2005.

M. Wuthrich, M. Liwicki, A. Fischer, E. Indermuhle, H. Bunke, G. Viehhauser, and M. Stolz. Language model integration for the recognition of handwritten medieval documents. In *Proc. of the 10th Int. onf. on Document Analysis and Recognition (ICDAR 2009)*, pages 211 –215, 2009.

CHAPTER 8

Scientific Contributions

The objective of this thesis was to study, develop and evaluate an interactive transcription platform for transcribing handwritten text document when user effort is limited. Summing up the contributions of this thesis are:

1. Development of a general interactive annotation approach The interactive annotation process has been presented from a theoretical point of view. In this approach, user interaction is added to the conventional classification problem. Due to the impossibility, in most applications, of a direct estimation in this approach, which would have to consider all possible interactions and system decisions, some assumptions can be performed. As result, two approaches have been presented for interactive transcription and document layout analysis of old text documents.

2. Acquisition of handwritten text databases Annotated documents are required in order to develop and assess new techniques in HTR. They also need to be freely available for external researchers to prove the effectiveness of the proposed methods and to help them to develop new techniques. In this thesis, we have collaborated in the annotation of two (old) handwritten text documents. These documents correspond to two old manuscripts from the XVI and the XVIII centuries. They were specially selected to cover all frequent problems in the transcription of old documents. We describe the digitisation and annotation procedure and present baseline HTR experiments to evaluate the difficulty of the task.

3. Development of interactive tools to transcribe documents This thesis deals with the development of new techniques to effectively transcribe documents in a CAT approach. To this purpose, we have developed an interactive prototype called GIDOC to deal with the interactive annotation of handwritten text documents. We have implemented and tested all the developed techniques in this thesis on this prototype. GIDOC can be used

without any external tools, and it covers all procedures require for HTR: document layout analysis, preprocessing, feature extraction, training and recognition. On document layout analysis, it includes tools to detect text blocks, and text baselines required in the next step. Several preprocessing techniques are included, such as noise removal and script normalisation. Training of HMMs and n-gram LMs is integrated in the prototype or can be performed by external software, HTK for HMMs, and SRILM for n-gram models. Again, recognition is included in the prototype along with hypothesis verification using CMs. The GIDOC prototype is freely available under GNU GPL3 license.

4. Development of methods to efficient interactive annotation of handwritten documents We have developed a series of methods and techniques to efficiently employ the user supervision available. Concretely, we focus on the CAT of handwritten text documents when user supervision available is limited. We have developed three main improvements in this approach. First, we efficiently employ user supervision by guiding it towards incorrectly recognised words by means of CMs. Second, user supervised words are used as constrains to recompute the current system hypothesis. User supervised words affect the previous system recognition as they reduce the uncertainty of the system. Third, at the end of the previous steps, both supervised and unsupervised transcriptions are used to improve the system.

5. Creation of methods to balance the error and user effort We have created an approach in which the CAT system objective is to reach a predefined error rate with the minimum user supervision effort possible. As a result, we have developed methods to estimate the expected error rate of the system output. This prediction is then used to supervise the output accordingly using the minimum effort possible and guaranteeing that the error rate is below the threshold predefined by the user. We also integrate this approach into GIDOC in order to efficiently employ the supervision.

APPENDIX A

The GIDOC Prototype

Contents

A.1 Introduction

As said in Chapter 2, due to the unsatisfactory results of current state-of-the-art systems, a better approach is to follow a CAT process. In this approach, the transcription task is completed by an user, which is continuously aided by a system reacting and learning from the interaction. However, the implementation of the described approach is not straightforward. On one hand, it requires the implementation of a whole HTR process, which comprises the use of several techniques and methods. First, layout analysis have to be applied to locate which parts of the image contains text blocks. Second, line segmentation algorithms detect the position of the lines within the image text blocks. Third, each line image is preprocessed to reduce the variability of the script. Finally, a HTR system would be trained and a used to annotate the document. On the other hand, user interaction with the system should be comfortable and friendly to allow efficient image annotation. Finally, final users, which will be mainly paleography experts, need transcription tools that free them from the details of the underlying system and help them to reduce the effort needed to transcribe documents.

In this appendix, a CAT system prototype is presented for handwritten text in old documents, which implements most of techniques and methods developed in this thesis. It is a first attempt to provide integrated support for interactive page layout analysis, text line detection and handwritten text transcription. Clearly, it is a programming challenge to develop a usable, friendly GUI for such a prototype, and thus we decided not to start from scratch, but to build it on top of the well-known GNU Image Manipulation Program (GIMP) (GIMP). Apart from its high-end user interface, GIMP gives us for free many desired prototype features such as 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. Indeed, the prototype, which will be referred to as GIDOC (Gimp-based Interactive transcription of old text DOCuments), is implemented as a set of GIMP plug-ins. GIDOC has been successfully used in the annotation of different handwritten old text document, such as GERMANA (Pérez et al., 2009) and RODRIGO (Serrano et al., 2010).

This appendix is structured as follows. A brief description of the whole prototype is given in Section A.2. Then, each of the remaining sections is devoted to each of the necessary steps required to complete the interactive transcription task. First, the preferences options of GIDOC are reviewed in Section A.3. Next, the block detection algorithms implemented are described in Sec A.4. Section A.5 explains how line detection is performed in GIDOC. Then, the preprocessing algorithms included in the implementation are introduced in Section A.6. The different feature extraction methods available in GIDOC are detailed in Section A.7. The HTR system training within GIDOC is reviewed in Section A.8. Next, Section A.9 depicts the GIDOC transcription dialog and its functionality. Finally, in Section A.10, conclusions are drawn and future work is analysed.

A.2 System Overview

As indicated by its name, GIDOC has been implemented on top of the well-known GNU Image Manipulation Program (GIMP). As GIMP, GIDOC is licensed under the GNU General Public License, and it can be freely downloaded from (GIDOC). In order to use GIDOC, we

must first run GIMP and open a document image, convert it to grayscale and save it in the XCF[a] format. XCF format is the image native format of GIMP, which stores the image, layers and other additional information required for GIDOC. Then, GIMP will come up with its high-end user interface, which is often configured to only show the main toolbox (with docked dialogs) and an image window. GIDOC can be accessed from the menubar of the image window (see Figure A.1).

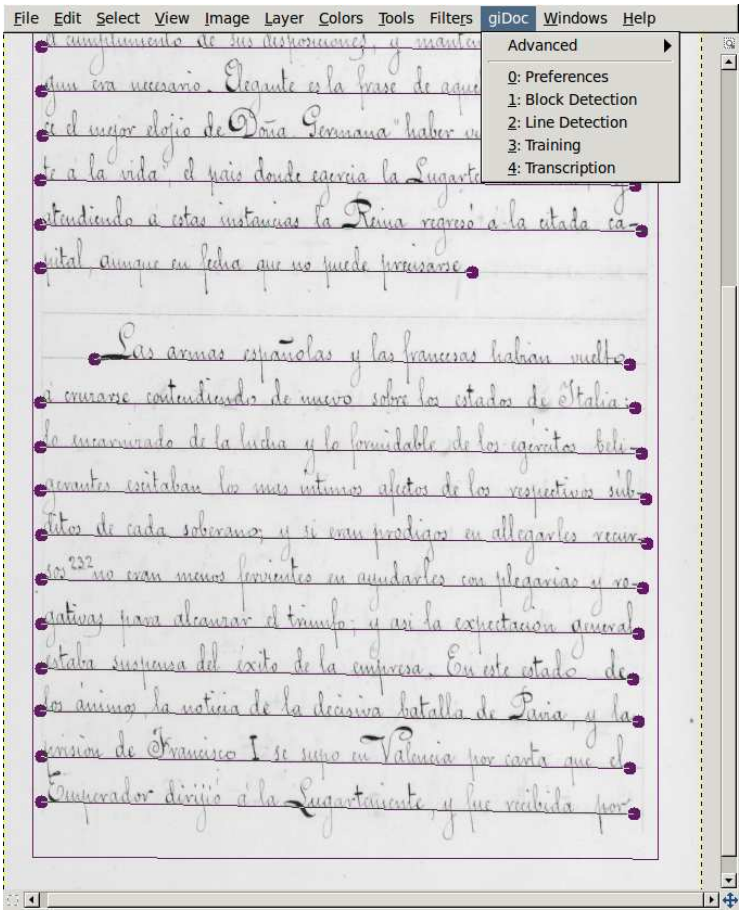


Figure A.1: Image window showing GIDOC menu.

As shown in Figure A.1, GIDOC menu includes six entries. First, *Advanced* options, where all atomic operations of the interactive HTR process can be applied individually. Next options list the operations that need to be followed to annotate an image. Each entry is formed by a number indicating the order of the step and its name. *0: Preferences*, in which global options, such as the project name, can be specified along with specific option for each part of

[a]More details about this file format in `http://en.wikipedia.org/wiki/XCF_(file_format)`

the process. *1: Block Detection* deals with the detection of text blocks in the image. *2: Line Detection* marks the text baseline of detected text blocks. *3: Training* builds the HTR system from all transcribed images. It must be noted that, this step has to be skipped at the beginning of the document transcription until the some pages have been annotated. *4: Transcription* option opens the transcription dialog, in which interactive transcription process is performed.

A.3 Preferences

GIDOC has been developed to assist transcribers in the transcription of different documents. As each of this documents possess different characteristics, GIDOC has been designed to manage them as different projects. Project files are stored inside the user home directory in a hidden folder called gidoc, i.e. *$HOME/.gidoc*, which includes the project configuration file along with the HTR system models. When a project is created or selected by the user it is marked as the active project. The active project configuration variables and its HTR system will be used by default when applying the rest of the tools.

The preferences option in the GIDOC menu opens a dialog, in which all the options and parameters of the interactive transcription process can be managed. If an active project is present, the preference dialog loads all its configuration variables. On the contrary, a new default project called "Germana" is created and its configuration files are set to their default values for the transcription of the GERMANA database, which is the database that was used as the benchmark of the prototype. Figure A.2 shows the main tab of the preferences dialog of the default project. At the top of the dialog, there are two buttons to create a new project (the left icon), and to open an existing project (the right icon). Below these buttons there is a tab menu, in which each option represent a different part of the transcription process. Each part will be described in its corresponding section in the following. At the center of the widget, project-related variables are shown. The name of the project is chosen upon creation and cannot be changed afterwards. Document directory defines a folder, in which the image files in XCF format of the documents are stored. This folder is used by training and recognition tools in order to locate the document files. The last variable, the "Lock Transcriptions" checkbox, disable the modification of stored transcriptions in document images.

A.4 Block Detection

Document layout analysis is a research field that deals with the identification and classification of logical entities residing in an image. For instance, the detection of the position of text blocks on a given image. During its development, GIDOC has been mainly tested on documents, in which most pages only contain nearly calligraphed text written on ruled sheets of well-separated lines, as in the example shown in Figure A.1. Consequently, GIDOC is designed to detect the text blocks in such documents taking advantage of their homogeneity. For all the block detection methods implemented, first, GIDOC performs an automatic detection of the text blocks within the image and then, the user revise and correct, if needed, the result. Text blocks are stored as GIMP paths and can be easily modified with the GIMP interface.

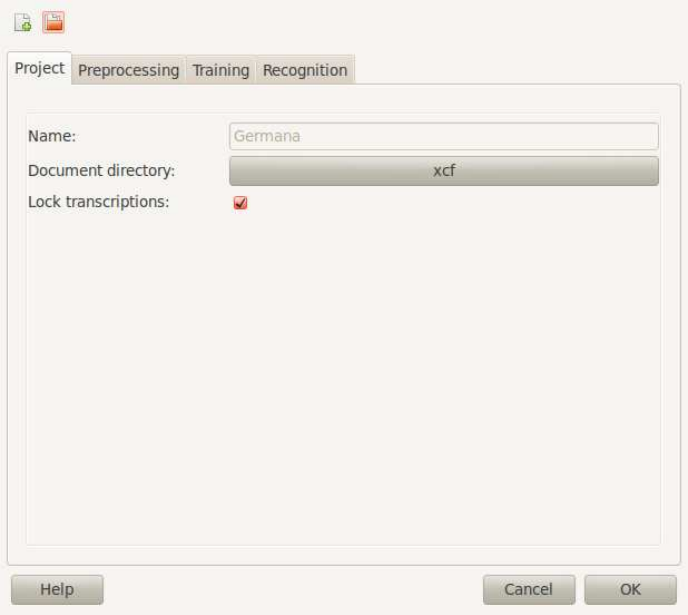


Figure A.2: Preferences dialog on GIDOC

A.4.1 Projection-based Block Detection

Typically, old text documents follow a well defined template, which should make easier the detection of its structure. However, document structure changes from one document to another and block detection techniques have to be adjusted independently on each of them. A standard and successful method to detect a great variety of document layouts is based on projection methods (Likforman-Sulem et al., 2007). A projection is the visual representation of the number of times a certain event occurs in each row or column of pixels in the image. For instance, if we represent the counts of each row as a histogram, we obtain the vertical histogram because it is represented vertically along the image. Similarly, if we perform the same process for each column, we obtain the horizontal histogram. In block detection, we are interested in a specific type of event, that is, the pixel value. In this case, the projection will correspond to the summation of the values of all pixels. In the following, these projections pixels are referred as *pixel value* projections.

However, projection-based methods need that text blocks in the images are totally straight, as a slight rotation modify the projection obtained. In rotated images, a previous step to block detection is needed in order to to reduce the global inclination of the image. This inclination is commonly known as Skew. Skew can be measures as the mean angle of the document. GIDOC implements the skew correction method developed in (Pastor, 2007), which detects the skew angle of a pages using vertical projections. Once the skew angle has been detected, GIDOC uses GIMP tools “undo” the rotation of the image.

As said, horizontal and vertical pixel value projections sum the value of a column, or

row of pixels, respectively. Given an image, its corresponding text blocks can be located by exploring the vertical and horizontal projections. For instance, Figure A.3 shows a page of the GERMANA database (Pérez et al., 2009) along with its vertical and horizontal projections. Note that, projections have been shadowed to not cover the image. This manuscript follows a fixed template, in which only a lone text block appears. As observed, it is easy to locate the text block coordinates by using the projections. First, the left border of the text block can be located by detecting in which pixel there is the maximum change in pixel values between itself and pixels nearby. Similarly, the right border can be located with the opposite process. Second, the top border of the text block can be located by a similar process. However, instead of selecting the maximum change, the process would rather locate the n-maximum and then select the one at the highest position. Again, the bottom border can be detected with the opposite process.

Even though the simplicity of the described process, it suffers from a strong dependence on the previous preprocess and other parameters. For instance, the summation of pixel values is directly affected by noise in the image, and thus, noise removal techniques have to be applied. These parameters change from one document to another, and have to be tuned specifically for the task. In addition, each different document layout will require a different process. The described process will only manage to detect blocks in GERMANA, or similar documents, in which a lone text block appears.

A.4.2 History-based Block Detection

GIDOC also implements a novel text block detection method, in which conventional, memoryless techniques are improved with a "history" model of text block positions. Typically, conventional block detection methods only consider information from current document image. In document collections with an homogeneous structure, a better approach is to include information of previously detected blocks when detecting the current one. For instance, in GERMANA, a lone text block appears in all document images, and its position is mainly located in two different positions. These two different positions depend on which side of the book was placed the page when the document was written. Right placed images are referred as "front" and left placed images are referred as "back". For instance, Figure A.4 shows a back page along with its following front page.

In (Ramos-Terrades et al., 2010), we considered the detection of text blocks in GERMANA as a classification problem. Each document image is classified into class as "front" or "back". We suppose that classification is part of a sequential process, in which document images are classified one after the other. In this process, it is assumed that the user has corrected the position and classes of all previously processed pages. Then, given the current document image, and the previous pages document classes and user feedback, the method presented in Section 3.4 is employed to obtain the document class and structure of the current page.

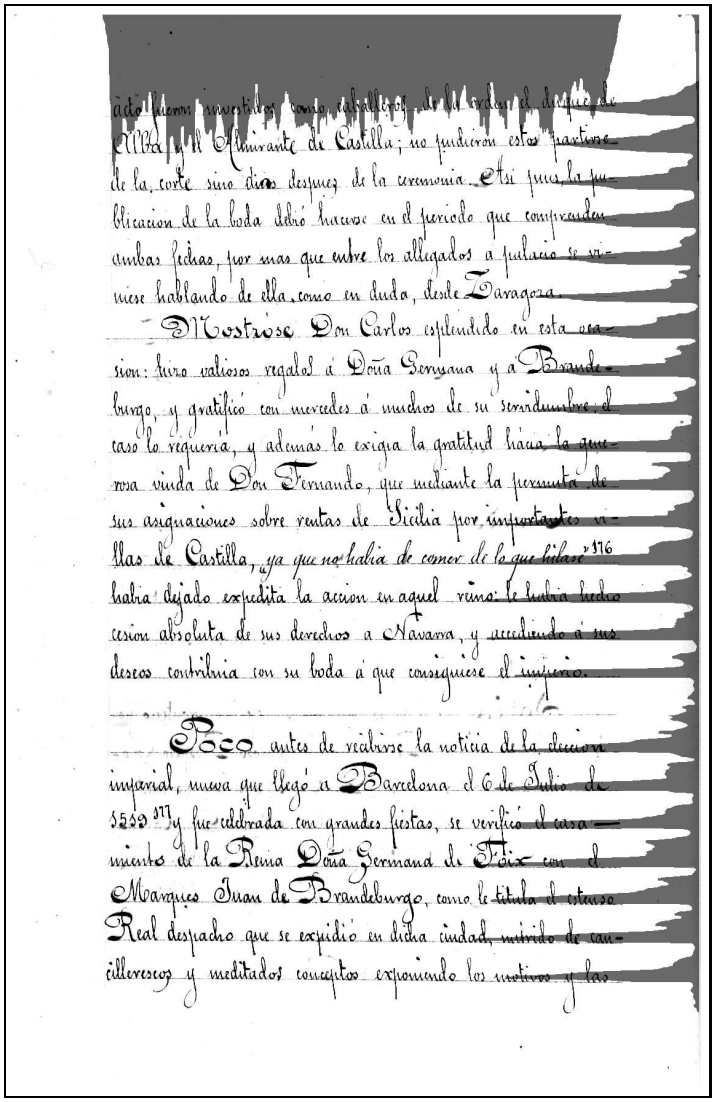


Figure A.3: Vertical and Horizontal pixel value projections of a GERMANA page.

A.5 Line Detection

Given a textual block, the *Line Detection* entry in the GIDOC menu detects all its text baselines, which are marked as straight paths using the path tool of GIMP. These paths can be easily adjusted with GIMP interface in those cases the automatic detection do not work. Line detection in GIDOC is also based on projection based methods, and it is performed in a similar way to the projection-based block detection method. However, in order to detect line

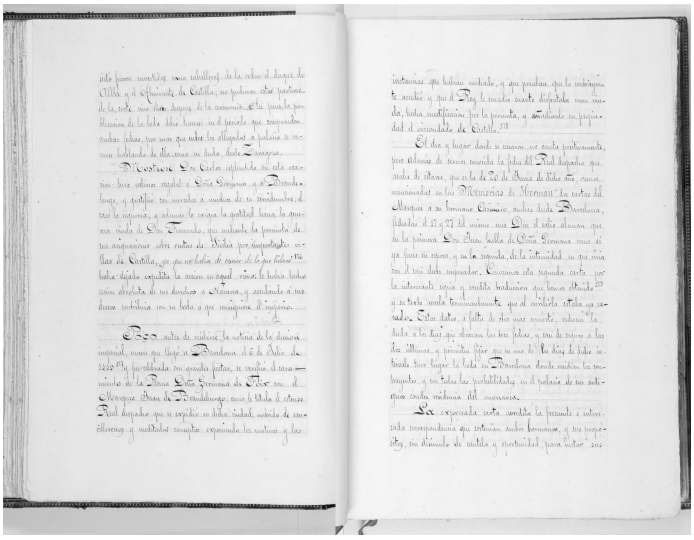


Figure A.4: Two consecutive pages of GERMANA.

baselines of an image, it is better to account for the number of black to white transitions instead of pixel values. The main reason for this change is that, for each row of pixels in the image, the number of black to white transitions is expected to be higher for rows containing handwritten letters, and lower for rows between the lines. Figure A.5 depicts a page of the GERMANA database, in which its vertical transition projection is estimated and depicted at the right side. As observed, line baselines can be located by exploring the projection. Concretely, GIDOC detects baselines by exploring which pixel column in the projection contains the maximum number of changes from black to white occur. If two columns posses the same number of transitions, GIDOC selects the closest to the right side. This process is similar to computing a horizontal projection of the vertical projection.

Even though the presented process correctly detects most of the lines, it does not work for short lines, in which the number of black to white transitions may not be high enough. These lines can be detected by finding wide gaps between detected lines, which are big enough to contain an undetected line. As example of this problem can be observed in the fourth line in Figure A.5. In GIDOC this process is refined by letting the user define the number of lines in the document. Old text documents typically follow a template and the number of lines remain unchanged in the whole manuscript. This refinement helps GIDOC to locate lines that have been undetected in the projection. The number of lines can be adjusted in the the preprocess tab in the preferences dialog.

The described line detection method manages to detect straight baselines within the document image. However, digitisation of old text documents may cause the lines to suffer from a slight warping distortion at one of its sides. This warping is caused by the binding of the document, which curves the baseline towards the center of the opened book when scanning the document. Currently, GIDOC do not include any tool to correct this warping, and it has

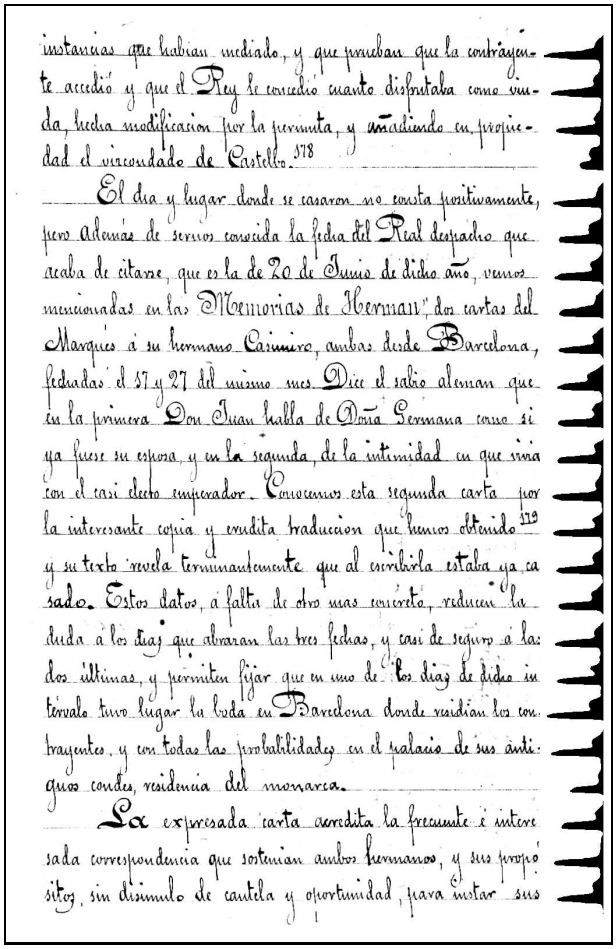


Figure A.5: Vertical projection counting the number of black to white transitions on a page in GERMANA.

to be corrected manually. In order to correct it, the user should defined a curved baseline with GIMP path tools. As previously happened with block detection, this step of the system will be always supervised by the user.

A.6 Preprocessing

Once document baselines have been marked, GIDOC extracts a text line image for each of them. For each point within the baseline, GIDOC extracts a perpendicular line of pixels in a similar way to a sliding window extraction approach. The number of pixels extracted along the baseline can be adjusted in the preprocess tab at the preferences dialog. Concretely, the

size of the extracted column is defined by two variables, under and over, which refers to the number of pixels extracted below and above the baseline. This procedure is able to extract a straight line image even when the annotated baseline follows a curved path.

Line images constitute the input of HTR systems. However, the use of raw images as an input leads to poor recognition results. Raw images contain a lot of noise due to the state of the document or the digitisation process applied. Furthermore, image character models of HTR systems are typically trained by sequentially processing each column of the image. This restriction requires that the slant of the script is corrected, in order to correctly train the image characters models. Finally, some characters of handwritten text lines posses long ascendants and descendants, such as the "t" or "l" letters. This characteristic difficulties the HTR process because the size of all letters is not uniform. A better approach is to apply a size normalisation process to make all letters fill the same dimensions, in which case, it is easier to discriminate between them. Figure A.6 depicts the described preprocessing process when applied to a line in GERMANA. In this document, first, noise removal processes are applied, second, slant is corrected and finally ascendants/decedents are normalised.

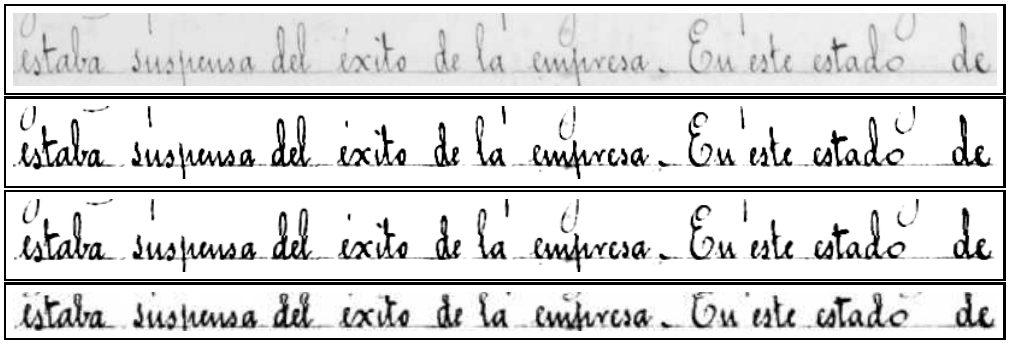


Figure A.6: Preprocessing of a text line image. From top to bottom: original image, denoising, deslanting and vertical size normalisation.

In practice, each document require different preprocessing steps, and they have to be tuned accordingly. GIDOC implements a wide variety of preprocessing tools, such as, median filter noise removal, or word slope correction. These tools are an adaptation of all preprocessing tools implemented by the HTR division inside the "Pattern Recognition and Human Language Technology" research group. An overall description of these methods can be found in (Pastor, 2007). In addition to these tools, GIMP also include many built-in tools for photo manipulation, which could be used if necessary. The preprocessing steps required for each document can also be adjusted in the preprocess tab at the preferences dialog (see Figure A.9) by defining a GIMP custom procedure in the entry line called "Custom Procedure". A GIMP custom procedure is a script, in which the preprocessing step are listed.

A.7 Feature Extraction

The preprocessing of raw images improves the result of HTR systems, eliminating the variability of the images. However, there is still much redundant information in preprocessed images, in which not all pixels in the image provide the same information. In PR tasks, objects are represented as features. These features are used to classify an object into a class. A feature discriminates better between classes if it posses a great variability. A feature in which no variability is observed, is not suitable for classification. In case of HTR, given a line image, the center part of each character varies more than the rest, and it is typically a better feature than the extracted from border pixels. In addition, individual features can be combined in a larger dimensional space, in which is easier to classify the object. As said in Section 2.3, the process of selecting a better representation of a given object is known as "Feature Extraction".

GIDOC implements two different feature extraction methods for HTR. The first one is motivated by the feature extraction used in ASR tasks. In this method, each pixel is processed according to its surrounding pixels, which defines a window within the image. For each pixel in the window, three different values are computed. First, the mean value of a gaussian modelling the window. Second and third, the horizontal and vertical derivative of the window. A detailed explanation of the process can be found in (Toselli et al., 2008). Figure A.7 shows an example of the described feature extraction method applied to the previously preprocessed example in Figure A.6. As observed, a higher quantity of information, compared to the preprocessed image, is contained in this new representation.

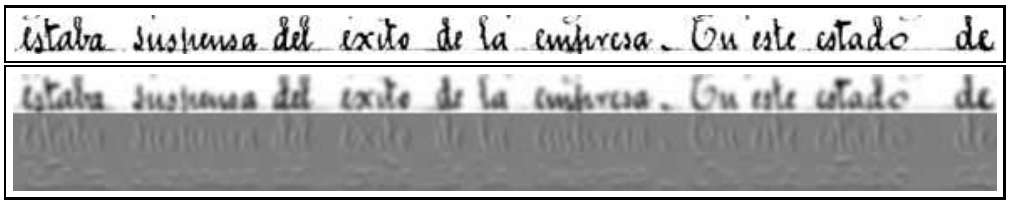


Figure A.7: Feature extraction of a text line image, From top to bottom: vertical size normalisation and a derivative-based feature extraction.

The other method included in GIDOC is an implementation of the feature extraction method used by the FKI research group in most of their papers (Marti and Bunke, 2002a). These feature extraction method extracts nine features for each column of the image. These correspond to some geometrically motivated features, such as the total pixel mass in the column, or the gravity center. These features were selected from a huge pool of other features while optimising the recognition of the standard IAM database (Marti and Bunke, 2002b). Figure A.8 shows the result of normalising the feature extraction obtained to a gray value in order to represent it.

Figure A.9 shows the preprocess tab in the preferences dialog, in which the feature extraction method used can be selected. As observed, this dialog also includes the variables and options of the block and line detection, along with the preprocess module.

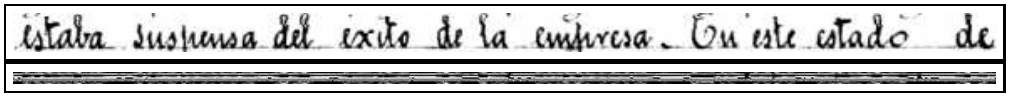


Figure A.8: Feature extraction of a text line image, From top to bottom: vertical size normalisation and geometrically motivated feature extraction.

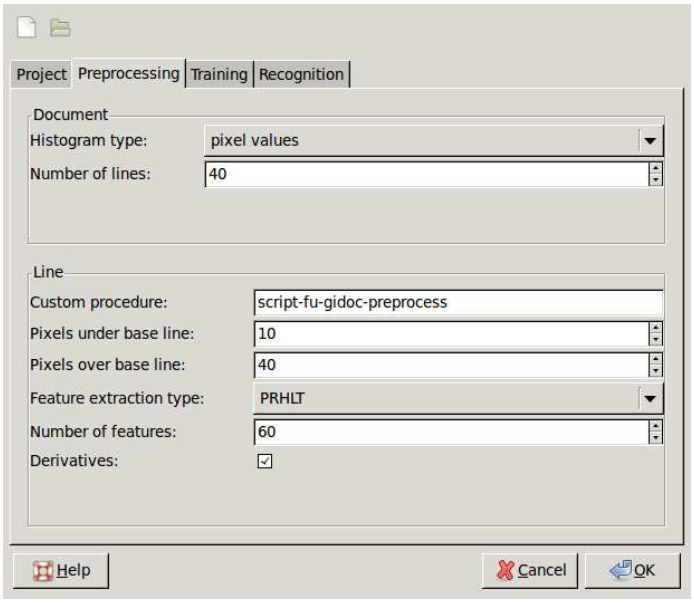


Figure A.9: Preprocess tab at the preferences dialog on GIDOC

A.8 Training

The *Training* option of GIDOC menu trains an HTR system from the documents images in the document project directory. GIDOC reads the directory of task document images and, for each image, it extracts all its transcribed text lines, if any, together with their corresponding line images. Then, each extracted line image is preprocessed following the user defined script and the selected feature extraction method is applied. Alternatively, transcriptions are first preprocessed to isolate special characters (mainly punctuation signs) and expand abbreviations. For instance *S.M.* is expanded to *Su Magestad*. Finally, the two parts of the HTR system are trained: the character image models, and the language models.

Image character models can be trained by two different toolkits: the standard HTK system (Young et al., 1995) or a GIDOC built-in toolkit, which is a standalone version of the AK toolkit (Giménez, 2011). Both of them train a HMM of a fixed number of states for each character of the extracted transcriptions, in which each state emits a gaussian mixture model. The training of the HMMs is performed by iteratively doubling the number of components of the mixture after a number of iterations of the EM algorithm has been applied. By default,

GIDOC trains an HMM with four states per character and 64 components per mixture, in which at each step of the estimation four iterations of the EM were applied. On the other hand, language models can be trained by two different toolkits. The SRI language model toolkit (Stolcke, 2002) or a built-in toolkit. GIDOC, by default, generates a bigram language model with Knesser-Ney discounting from the extracted annotations.

All the defined training parameters can be manually adjusted in the train tab at the preferences dialog (see Figure A.10), along with the toolkit used for HMMs estimation or the command line applied to train language models. The trained HTR models are stored in the project directory. It must be noted that, the execution of this module may be time-consuming and it is expected to be applied by transcribers seldomly, when sufficiently significant new data have been annotated.

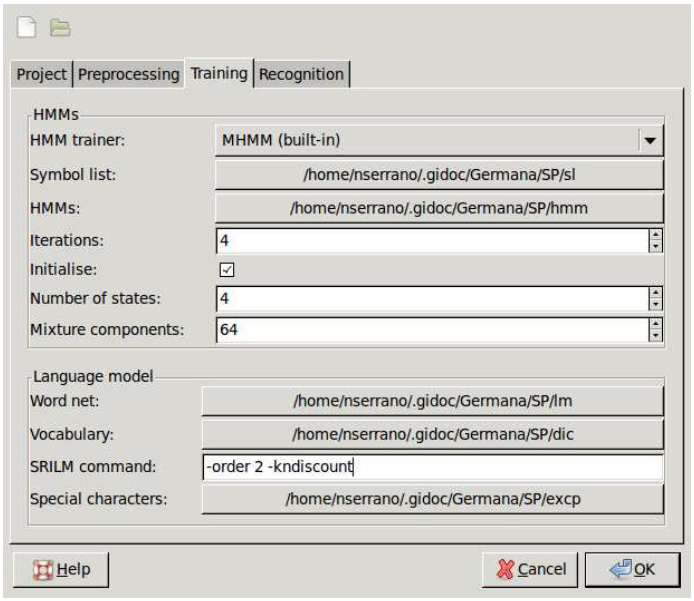


Figure A.10: Train options in the preferences dialog of GIDOC

A.9 Transcription

The *Transcription* entry in the GIDOC menu opens the interactive transcription dialog (see Figure A.11). It consists of two main sections: the image section, in the upper part, and the transcription section, in the bottom part. A number of text line images are displayed in the image section together with their transcriptions, if available, in separate editable text boxes within the transcription section. The *current* line to be transcribed or simply supervised is selected by placing the edit cursor in the appropriate editable box. Its corresponding baseline is emphasised (in blue color) and, whenever possible, GIDOC shifts line images and their

transcriptions so as to display the current line in the central part of both the image and transcription sections. It is expected that the user transcribes or supervises text lines, from top to bottom. However a different order can be followed, by entering text and moving the edit cursor with the arrow keys or the mouse.

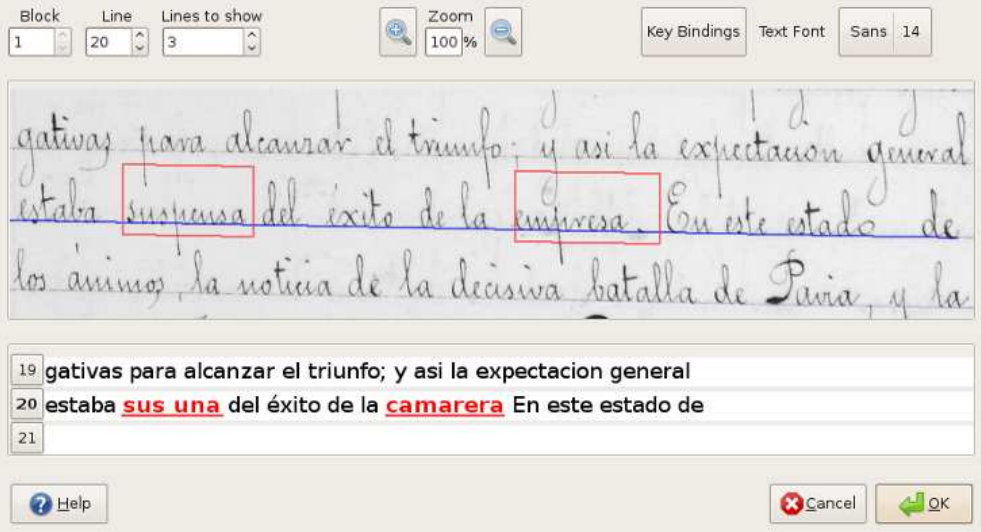


Figure A.11: Interactive transcription dialog.

As seen in Figure A.11, each editable text box in the transcription section, has a button attached to its left. This button is labelled with the corresponding line number. By clicking on it, its associated line image is extracted, preprocessed, transformed into a sequence of feature vectors, Viterbi-decoded using HTK and the models in the *Training* phase, and confidences measures are extracted. Recognition parameters are defined in the recognition tab at the preference dialog (see Figure A.12). Then, each recognised word is highlighted in red in both, transcription and text image, if confidence measure is below a defined threshold. In this way, it is not needed to enter the complete transcription of the current line, but hopefully only minor corrections to the decoded output. Clearly, this is only possible if, first, text lines are correctly detected and, second, the HMM and language models are adequately trained, from a sufficiently large amount of training data. Therefore, it is assumed that small quantity of transcription are manually annotated to train a preliminary HTR system.

A.10 Conclusions & Future Work

A computer-assisted transcription prototype called GIDOC has been presented for handwritten text in old documents. GIDOC is a first attempt to provide integrated support for interactive page layout analysis, text line detection and handwritten text transcription. It is build on top of GIMP, and uses standard techniques and tools for handwritten text preprocessing and

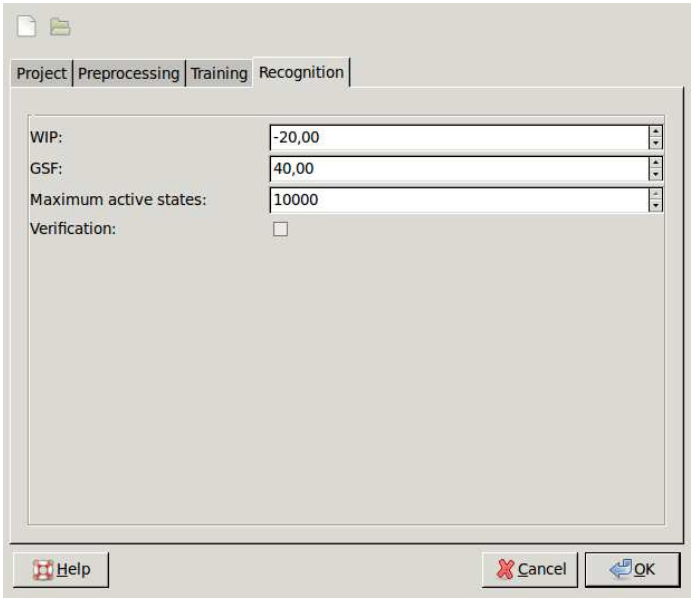


Figure A.12: Recognition tab at the Preferences dialog.

feature extraction, HMM-based image modelling, and language modelling. As GIMP, GIDOC is licensed under GNU General Public License, and it can be freely downloaded from Internet. The effectiveness of GIDOC has been empirically demonstrated on GERMANA, RODRIGO and ESPOSALLES databases.

Even though this prototype has been shown to be effective in the annotation of real documents, the prototype is at the beginning stages of developments and much work still needs to be done so that the prototype can be used by the general public. The prototype installation is not straightforward and this difficulty may discourage the user. Real user feedback is needed to ensure the quality and usability of the tools. Similarly, there are additional difficulties due to peculiarities in some documents, such as multilinguality, that need to be considered by GIDOC. However, the current prototype version is hard to upgrade and modify. It remains as a future work, the implementation of the prototype as a library of stand alone functions and the development of a complete manual and API.

The prototype has been presented in an international workshop:

- **N. Serrano** and L. Tarazón and D. Pérez and O. Ramos-Terrades and A. Juan. The GIDOC prototype. In *Proceedings of the 10th International Workshop on Pattern Recognition in Information Systems (PRIS 2010)* , Funchal (Portugal) June 2010.

Bibliography

GIDOC. Gimp-based Interactive transcription of old text DOCuments. Please visit `http://sourceforge.net/projects/gidoc/`.

A. Giménez. Adria's kit. `http://aktoolkit.sourceforge.net`, 2011.

GIMP. Gnu image manipulation program. `http://www.gimp.org/`.

L. Likforman-Sulem, A. Zahour, and B. Taconet. Text line segmentation of historical documents: a survey. *Int. Journal on Document Analysis and Recognition (IJDAR)*, 9, 2007.

U.-V. Marti and H. Bunke. In *Hidden Markov models*, chapter Using a statistical language model to improve the performance of an HMM-based cursive handwriting recognition systems, pages 65–90. 2002a.

U. V. Marti and H. Bunke. The IAM-database: an English sentence database for off-line handwriting recognition. *International Journal on Document Analysis and Recognition (IJDAR)*, pages 39–46, 2002b.

M. Pastor. *Aportaciones al reconocimiento automático de texto manuscrito*. PhD thesis, Dep. de Sistemes Informàtics i Computació, 2007.

D. Pérez, L. Tarazón, N. Serrano, O. Ramos-Terrades, and A. Juan. The GERMANA database. In *Proc. of the 10th Int. Conf. on Document Analysis and Recognition (ICDAR 2009)*, pages 301–305, Barcelona (Spain), 2009.

O. Ramos-Terrades, N. Serrano, A. Gordó, E. Valveny, and A. Juan. Interactive-predictive detection of handwritten text blocks. In *Proc. of SPIE-IS&T Electronic Imaging (DDR XVII)*, pages 75340Q–(1–10), 2010.

N. Serrano, F. Castro, and A. Juan. The RODRIGO database. In *Proc. of the 7th Int. Conf. on Language Resources and Evaluation (LREC 2010)*, pages 2709–2712, 2010.

A. Stolcke. SRILM - An Extensible Language Modeling Toolkit. In *Proc. of the Int. Conf. on Spoken Language Processing*, pages 901–904, 2002.

A. H. Toselli, V. Romero, M. P. i Gadea, and E. Vidal. Preprocessing and feature extraction technics for multimodal interactive transcription of text images. Technical report, Instituto Tecnológico de Informàtica, 2008.

S. Young et al. *The HTK Book*. Cambridge University Engineering Department, 1995.

List of Figures

List of Tables