Computer Vision and Image Understanding
journal homepage: www.elsevier.com

Image-speech combination for interactive computer assisted transcription of handwritten documents

Emilio Granell, Verónica Romero, Carlos-D. Martínez-Hinarejos**

PRHLT Research Center, Universitat Politècnica de València, Camino de Vera, s/n - 46022 Valencia - Spain

ABSTRACT

Handwritten document transcription aims to obtain the contents of a document to provide efficient information access to, among other, digitised historical documents. The increasing number of historical documents published by libraries and archives makes this an important task. In this context, the use of image processing and understanding techniques in conjunction with assistive technologies reduces the time and human effort required for obtaining the final perfect transcription. The assistive transcription system proposes a hypothesis, usually derived from a recognition process of the handwritten text image. Then, the professional transcriber feedback can be used to obtain an improved hypothesis and speed-up the final transcription. In this framework, a speech signal corresponding to the dictation of the handwritten text can be used as an additional source of information. This multimodal approach, that combines the image of the handwritten text with the speech of the dictation of its contents, could make better the hypotheses (initial and improved) offered to the transcriber. In this paper we study the feasibility of a multimodal interactive transcription system for an assistive paradigm known as Computer Assisted Transcription of Text Images. Different techniques are tested for obtaining the multimodal combination in this framework. The use of the proposed multimodal approach reveals a significant reduction of transcription effort with some multimodal combination techniques, allowing for a faster transcription process.

1. Introduction

Transcription of handwritten documents is an interesting application field where converge the use of image processing (for image feature extraction) and natural language processing (for text recognition). This application can be used to obtain a text transcription of handwritten notes and documents, such as historical handwritten books; this last task is specially interesting for the preservation of cultural heritage available in different libraries (Fischer et al., 2009). In recent years, Handwritten Text Recognition (HTR) systems (Romero et al., 2012) have contributed to speed up the transcription of these manuscripts.

HTR systems differ from Optical Character Recognition (OCR) systems or scene text recognition systems in different features. OCR and scene text recognition deal with separated characters of different regular fonts, which makes possible to easily isolate each character and classify it by using this segmentation and, possibly, its context (Bissacco et al., 2013). However, in HTR it is difficult to make character segmentation and the written sequence must be taken as a whole, and the characters in cursive style present a much higher variability (even for the same writter). Moreover, OCR and scene text recognition usually deal with the recognition of single words or short word sequences (Jaderberg et al., 2014a), whereas HTR deals with longer sequences (usually in the form of lines of handwritten text). In contrast, text detection for HTR is usually easier than for scene text, because of the regular nature of most handwritten documents that made them easy to segment text blocks into lines.

HTR systems provide a draft transcription that human users can amend with less effort than transcribing from scratch. In these terms, HTR systems can be seen as a part of an assistive technology. Assistive technologies have been of traditional use in many fields of computer applications, such as the Computer Aided Design (CAD) field (Machover, 1995), medical diagnosis (Doi, 2007), automatic driving (Malit, 2009), Computational Linguistics/Natural Language Processing (CL/NLP) (Barrachina et al., 2009; Revuelta-Martínez et al., 2012; Silvestre-Cerdà et al., 2013), and Pattern Recognition and Image Processing (Romero et al., 2012).

In these tasks, the computer allows the human user to have

**Corresponding author:
e-mail: cmartine@dsic.upv.es (Carlos-D. Martínez-Hinarejos)

an easier and faster work, providing the final user with a series of tools that allow to speed-up the process. Among these tools appear some automatic processing elements, such as medical data analysers, processors for data from driving sensors, speech and handwritten text recognisers, image feature extractors, etc. Apart from that, these systems need an interface that allows the user to amend the possible errors obtained by the automatic process. The interface tool could provide, by using underlying systems that employ the results of the automatic process and the user feedback, autonomous actions that avoid the user to perform some of the corrections.

Consequently, the main objective in these systems is not obtaining the most accurate result from the automatic system, but achieving the lowest effort for the human user (although both facts could be correlated). This requires new evaluation measures and frameworks that follow this criterion (minimising user effort) and are adapted to the corresponding task. For example, for the transcription of speech or handwritten text, the number of correction actions that the user has to perform (taking into account automatic corrections given by the system) is a good measure of the effort.

In this assistive context, the multimodal paradigm arose as a new form of improving these systems by reducing the final user effort. The multimodal paradigm has experimented a spectacular growth in the latest years because of the development of mobile devices (Di Fabbrizio et al., 2009), where different modalities (speech and touch mainly) are employed for the device management. In the case of Image or Natural Language Processing tasks, multimodality has been applied to problems where signals of different nature that represent the same final object are available (Mihalcea, 2012; Potamianos et al., 2003; Sebe et al., 2005; Granell and Martínez-Hinarejos, 2015b). In any case, multimodality is strongly linked to human-computer interaction, since the user may employ different modalities to obtain a more ergonomic or faster interaction to achieve an objective.

One interesting computer assisted application where multimodality can provide productivity improvements is the transcription of handwritten documents (Gordo et al., 2008). In this case, the assistive system provides the final users an initial draft transcription of the handwritten image. Then, the system could supply with alternative transcriptions every time the user makes an amendment, with the final aim of reducing the user effort to obtain a perfect transcription.

An example of assistive framework that presents these features is the Computer Assisted Transcription of Text Images (CATTI) system (Romero et al., 2012). The CATTI system takes as input the image to be transcribed, which is employed to offer the user a first hypothesis and alternatives when the user makes a correction. The obtention of the hypotheses is usually based on a HTR system. Since HTR systems commonly take as input text lines, input images are usually text lines obtained from page images. The obtention of text lines from an initial page requires several steps that are common for degraded documents: slope correction (Bloomberg et al., 1995), bright normalisation, image cleaning (Villegas et al., 2015) and line segmentation (Grüning et al., 2018). Current state of the art for these methods provide high quality images of lines with a very accurate segmentation.

The multimodality can be incorporated into CATTI by providing another signal that represents the same sequence of words, e.g., a speech dictation of the text that can be processed by an Automatic Speech Recognition (ASR) engine and gives as a result different alternatives. HTR and ASR systems employ similar models: optical/acoustical models for the basic units (characters and phones, respectively), lexical models (to form words from basic units), and language models (to form sentences from words). These systems also can obtain results in a similar format: single best hypothesis, n-best list of hypotheses, or lattices to represent alternative hypotheses. Therefore, its combination seems feasible despite of the different nature of the signals and its asynchrony.

In this paper, we research the use of multimodal combination techniques for improving the performance of the CATTI framework, and in particular the interaction effort of the CATTI user. The baseline CATTI system employs HTR to obtain the initial draft transcription to be amended using the interactive protocols. We explore how the addition of a speech signal with the dictation of the handwritten text improves the CATTI performance by using multimodal combination of HTR and ASR results. Four different multimodal combination techniques are tested and compared with the use of a single modality in CATTI. Results show that multimodal combination can provide significant effort reductions for the final user.

Thus, the main contribution of the paper is to demonstrate that including multimodality (image and speech) as input for the CATTI system reduces substantially, by using proper multimodal combination techniques, the user effort when transcribing handwritten text lines.

The paper is organised as follows: Section 2 specifies the particulars of the new multimodal CATTI system; Section 3 presents the different multimodal combination techniques; Section 4 details the experimental framework (data, conditions, and assessment measures); Section 5 shows the results of the different experiments; Section 6 offers the final conclusions and future work lines.

2. Computer Assisted Transcription of Text Images

Many documents used every day include handwritten text and, in many cases, it would be interesting to recognise these text images automatically. However, state-of-the-art handwritten text recognition systems (HTR) can not suppress the need of human work when high quality transcriptions are needed. HTR systems can achieve fairly high accuracy for restricted applications with rather limited vocabulary (reading of postal addresses or bank checks) and/or form-constrained handwriting. However, in the case of unconstrained transcription applications, the current HTR technology typically only achieves results which do not meet the quality requirements of practical applications. Therefore, once the full recognition process of one document has finished, heavy human expert revision is required to really produce a transcription of standard quality. Such a post-editing solution is rather inefficient and uncomfortable for the human corrector.

A way of taking advantage of the HTR system is to combine it with the knowledge of a human transcriber, constituting the so-called Computer Assisted Transcription of Text Images (CATTI) scenario (Romero et al., 2012). As previously commented, in this framework the automatic HTR system and the human transcriber cooperate interactively to obtain the perfect transcript of the text images. At each interaction step, the system uses the text image and a previously validated part (prefix) of its transcription to propose an improved output. Then, the user finds and corrects the next system error, thereby providing a longer prefix which the system uses to suggest a new, hopefully better continuation.

Speech dictation of the handwritten text can be used as an additional information source in the CATTI process. Taking into account both the handwritten text image and the speech signal the system can, hopefully, propose a better transcription hypothesis in each interaction step. This way, many user corrections are avoided. In this section we review the classical HTR and ASR framework and formalise the multimodal CATTI scenario where both sources help each other to improve the system accuracy.

2.1. HTR and ASR Framework

The traditional HTR and ASR recognition problems aim to recover the text represented in an input signal, and therefore they can be formulated in a very similar way. However, the input signal for HTR systems usually is a segmented line of a digitalised handwritten document (Romero et al., 2015), whereas in ASR the input is a voice signal. Then, the problem is finding the most likely word sequence, $\hat{w}$, for a given handwritten text line image or a speech signal represented by a feature vector sequence $x = (x_1, x_2, \ldots, x_{|x|})$ (Toselli et al., 2004), that is:

$$\hat{w} = \arg\max_{w \in W} \mathrm{P}(w \mid x) = \arg\max_{w \in W} \mathrm{P}(x \mid w)\mathrm{P}(w) \tag{1}$$

where:

- W denotes the set of all permissible sentences,
- $\mathrm{P}(w)$ is the probability of $w = (w_1, w_2, \ldots, w_{|w|})$ approximated by the language model (Jelinek, 1998) (usually modelled by a n-gram word language model), and
- $\mathrm{P}(x \mid w)$ is the probability of observing x by assuming that w is the underlying word sequence for x, evaluated by the optical or acoustical models for HTR and ASR respectively (typically it is approximated by concatenated hidden Markov models -HMMs- that model the different characters or phonemes or by Deep Neural Networks -DNN- that model this probability distribution).

The search (or decoding) of $\hat{w}$ is carried out by the Viterbi algorithm (Jelinek, 1998). From this dynamic-programming decoding process, we can obtain not only a single best hypothesis, but also a huge set of best hypotheses. These solutions can be presented in the form of a n-best list or compactly represented into a lattice, such as a Word Graph (WG) or a Confusion Network (CN) (Jurafsky and Martin, 2009).

A WG is a weighted directed acyclic graph that represents a huge set of hypotheses in a very efficient way. It is defined as a finite set of nodes Q and edges E, including an initial node $v_I \in Q$ and a set of final nodes $F \subseteq (Q - v_I)$. Each node v is associated with a horizontal position for HTR or a time point for ASR of x, given by $t(v) \in [0, |x|]$, where $t(v_I) = 0$ and $\forall_{v_F \in F} t(v_F) = |x|$. Their edges are labelled with words and weighted with scores derived from the optical/acoutical and language model probabilities computed during the decoding process.

A *complete path* of a WG is a sequence of nodes starting with node v_I and ending with a node in F. Complete paths correspond to whole decoding hypotheses.

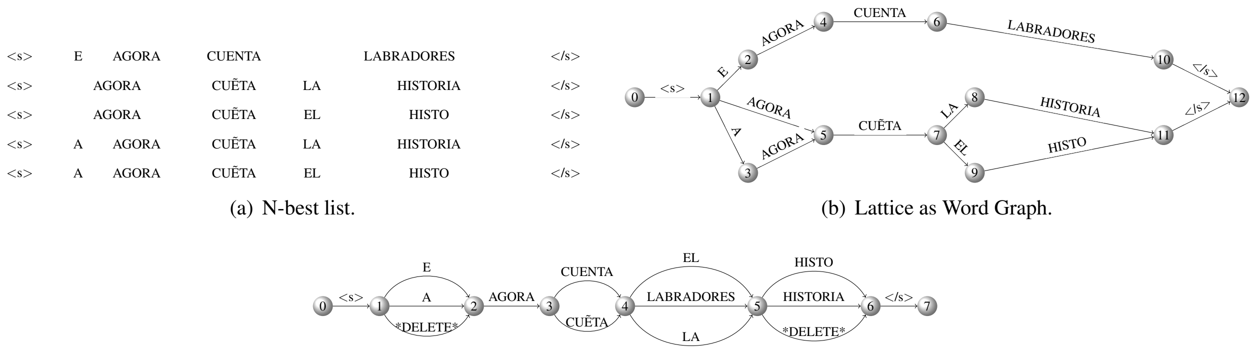
A CN is also a directed, acyclic and weighted graph that shows at each point which word hypotheses are competing or mistakable. Therefore, the segmentation information ($t(v)$ in WGs) is not available here. Each hypothesis goes through all the nodes by choosing one word from each position. To cope with different length hypotheses, the *DELETE* arcs are used. The words and their posterior probabilities are stored in the edges, and the total probability of the words contained in a subnetwork (all edges between two consecutive nodes) sum up to 1. It is important to note that confusion networks add paths that are not in the original recognition. Figure 1 provides an example of a n-best list, a lattice formatted as WG representing these n-best hypotheses, and an equivalent CN[1].

2.2. CATTI Formal Framework

As previously explained, in the CATTI framework the user is directly involved in the transcription process, since he/she is responsible for validating and/or correcting the system hypothesis during the transcription process. The system takes into account the handwritten text image and the feedback of the user in order to improve these proposed hypotheses. The more information the system has about what is written in the handwritten text line image, the better the proposed hypotheses are, and therefore, fewer user interactions are needed to obtain the perfect transcript. In this work, in addition to the handwritten text line image, we study how the CATTI system can take advantage of the speech dictation of the text that the images contain.

The process starts when the system proposes a full transcription $\hat{s}$ of the handwritten text line image. If, in addition to the handwritten text line image, the speech dictation is available, the system takes it into account to propose a hopefully better hypothesis. Then, the user reads this transcription until finding a mistake and makes a mouse action (MA) m, or equivalent pointer-positioning keystrokes, to position the cursor at this point. By doing so, the user is already providing some very useful information to the system: he is validating a prefix p of the transcription, which is error-free, and in addition, he is signalling that the following word e located after the cursor is incorrect. Hence, the system can already take advantage of this fact and directly propose a new suitable suffix (i.e., a new $\hat{s}$) in which the first word is different from the first wrong word of

[1] For the sake of simplicity, the probabilities have been omitted.



<s> E AGORA CUENTA LABRADORES </s>
<s> AGORA CUẼTA LA HISTORIA </s>
<s> AGORA CUẼTA EL HISTO </s>
<s> A AGORA CUẼTA LA HISTORIA </s>
<s> A AGORA CUẼTA EL HISTO </s>
0
<s>
1
E
2
AGORA
4
CUENTA
6
LABRADORES
10
</s>
12
AGORA
5
CUẼTA
7
LA
8
HISTORIA
11
</s>
A
3
AGORA
EL
9
HISTO
0
<s>
1
E
A
DELETE
2
AGORA
3
CUENTA
CUẼTA
4
EL
LABRADORES
LA
5
HISTO
HISTORIA
DELETE
6
</s>
7

(a) N-best list.

(b) Lattice as Word Graph.

(c) Lattice as Confusion Network.

Fig. 1. Output formats for recognition systems. The n-best list presents the *n* output hypotheses of higher probability obtained during the decoding process, ordered from higher to lower. Lattice representations provide the same information in a more compact format; in the case of Word Graph, segmentation is kept, whereas in Confusion Network it gets lost. The three representations in the figure are for the same output result. Probabilities are not shown for the sake of clarity.

the previous suffix. This way, many explicit user corrections are avoided (Romero et al., 2009). If the new suffix $\hat{s}$ corrects the erroneous word, a new cycle starts. However, if the new suffix has an error in the same position than the previous one, the user can make a new MA or can enter a word v to correct the erroneous one. This last action produces a new prefix p (the previously validated prefix followed by the new word v). Then, the system takes into account the new prefix to suggest a new suffix and a new cycle starts. This process is repeated until a correct transcription is accepted by the user.

In Figure 2 we can see an example of the CATTI process. In this example, without interaction with a CATTI system, a user should have to correct about three errors from the original recognised hypothesis (“abadia”, “segun” and “el”). Using CATTI only one explicit user-correction is necessary to get the final error-free transcription: the iteration 1 only needs a MA, but in the iteration 2 a single mouse action does not succeed and the correct word needs to be typed.

Formally, in the traditional CATTI framework (Romero et al., 2012), the system uses a given feature sequence, x_{htr}, representing a handwritten text line image and a user validated prefix p of the transcription. In this work, in addition to x_{htr}, a sequence of feature vectors x_{asr}, which represents the speech dictation of the handwritten text line image, is used to improve the system performance. Therefore, the CATTI system should try to complete the validated prefix by searching for a most likely suffix $\hat{s}$ taking into account both sequences of feature vectors:

$$\hat{s} = \arg\max_{s} \mathrm{P}(s \mid x_{htr}, x_{asr}, p) \tag{2}$$

Making the naive assumption that x_{htr} does not depend on x_{asr}, and applying the Bayes’ rule, we can rewrite Equation (2) as:

$$\hat{s} = \arg\max_{s} \mathrm{P}(x_{htr} \mid p, s) \cdot \mathrm{P}(x_{asr} \mid p, s) \cdot \mathrm{P}(s \mid p) \tag{3}$$

where the concatenation of p and s is w. As in conventional HTR and ASR, $\mathrm{P}(x_{htr} \mid p, s)$ and $\mathrm{P}(x_{asr} \mid p, s)$ can be approximated by HMMs or DNN, and $\mathrm{P}(s \mid p)$ by a language model (usually an n-gram) conditioned by p. Therefore, the search must be performed over all possible suffixes of p (Romero et al., 2012).

This suffix search can be efficiently carried out by using Word Graphs (WG) (Romero et al., 2012) or Confusion Networks (CN) (Granell et al., 2016) obtained from the combination of the HTR and ASR recognition outputs. In each interaction step, the decoder parses the validated prefix p over the WG or CN and then continues searching for a suffix which maximises the posterior probability according to Equation (3). This process is repeated until a complete and correct transcription of the input text line image is obtained. Therefore, the combination techniques applied on the HTR and ASR recognition results may have an impact on the interactive process. Section 3 describes different combination options for the two modalities.

3. Multimodal Combination

Multimodal combination is a problem that has been faced in different recognition systems. This section presents the different alternatives according the stage of the recognition system where the combination is performed (Section 3.1) and the specific techniques that are employed in this work (Section 3.2).

3.1. Combination alternatives

The combination of natural language recognition systems allows to improve the recognition accuracy. In most cases, this combination can be performed in three different stages of the recognition process (Li, 2005): in the feature extraction stage (feature combination), in the search process (probability combination), and in the decoding output (hypothesis combination).

- **Feature combination:** Feature combination is performed concatenating the different features at feature vector level to form a new feature vector sequence to be used in the recognition process (Potamianos and Neti, 2001). This combination method usually requires synchronous parallel feature streams.

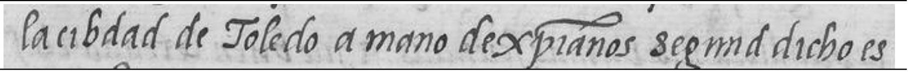
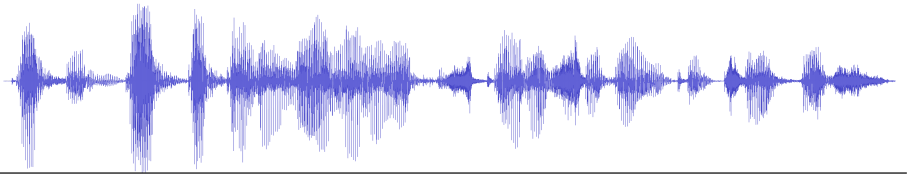
Image		
Speech		
ITER-0	p	
ITER-1	$\hat{s}$	*la abadia de Toledo a mano de xpiānos segun el dicho es*
	m	⇑
	p	*la*
ITER-2	$\hat{s}$	*cibdad de Toledo a mano de xpiānos segun el dicho es*
	m	⇑
	p	*la cibdad de Toledo a mano de xpiānos*
	- - - -	- -
	$\hat{s}$	*sigue el dicho es*
	v	***segund***
	p	*la cibdad de Toledo a mano de xpiānos segund*
FINAL	$\hat{s}$	*dicho es*
	v	#
	$p \equiv t$	*la cibdad de Toledo a mano de xpiānos* ***segund*** *dicho es*

Fig. 2. Example of CATTI operation using mouse-actions (MA). Starting with an initial recognised hypothesis $\hat{s}$ from combination of both modalities, the user validates its longest well-recognised prefix p, making a MA m, and the system emits a new recognised hypothesis $\hat{s}$. As the new hypothesis corrects the erroneous word, a new cycle starts. Now, the user validates the new longest prefix p, which is error-free, making another MA m. The system provides a new suffix $\hat{s}$ taking into account this information. As the new suffix does not correct the mistake, the user types the correct word v, generating a new validated prefix p. Taking into account the new prefix, the system suggests a new hypothesis $\hat{s}$. As the new hypothesis corrects the erroneous word, a new cycle starts. This process is repeated until the final error-free transcription t is obtained. The underlined boldface word in the final transcription is the only one which was corrected by the user. Note that in the iteration 2 it is needed two user interactions (a MA and then, to type the correct word). However, the iteration 1 only needs a user interaction (a MA).

- **Probability combination:** In probability combination methods the recognition class probabilities are combined before the final search process. The probability combination can be performed synchronously (Hernando et al., 1995) combining the observation probabilities of the optical/acoutical models frame-by-frame, or asynchronously (Dupont and Luettin, 2000) combining the probabilities at a higher-level, such as characters or phonemes. Synchronous probability combination requires synchronous parallel feature streams, while asynchronous probability combination allows to combine asynchronous parallel feature streams of the same nature (they use the same higher-level unit, such as in audio-visual speech recognition).

- **Hypothesis combination:** The last stage where the combination can be performed is at recogniser output (Fiscus, 1997). In this stage, the hypotheses obtained after the completion of the search process from each recogniser are combined. In hypothesis combination the parallel feature streams can be synchronous or asynchronous, and the only restriction is that all feature streams must represent the same final sequence of words.

Since HTR and ASR systems share most part of the recognition process, the possibility of combining both systems arises immediately. This combination would take advantage from two different data sources. However, this multimodal combination can not be performed easily at the input (feature combination) or during the recognition process (probability combination) given the different nature of these modalities and the asynchrony with respect to each other. Therefore, the easiest way of performing this multimodal combination is to combine the output results of both systems by using a hypothesis combination method, which is the option we selected for this work.

3.2. Hypothesis combination techniques

Many techniques on joining results have been proposed with the idea of reducing the error in the combined output. Some examples are: Recogniser Output Voting Error Reduction (ROVER) (Fiscus, 1997), N-best ROVER (Stolcke et al., 2000), Lattices Rescoring (Stolcke et al., 1997), and Confusion Network Combination (CNC) (Evermann and Woodland, 2000). These methods can be used to combine the outputs of recognition systems of different modalities that represent the same sentence. They all effectively improve the recognition performance, even though each one presents different characteristics.

3.2.1. Recogniser Output Voting Error Reduction (ROVER)

The widely used ROVER method (Fiscus, 1997) misses part of the information contained in the recognition outputs as it performs the combination by voting (at word level) among the different system outputs using only the 1-best hypothesis.

The ROVER method is implemented in two modules. In the first one, the 1-best decoding outputs are aligned and combined in a word transition network (with a structure similar to a confusion network). Then, the second module (the voting search module) evaluates each subnetwork to select the best scoring word (using a voting scheme) for the new transcription.

Voting is performed as follows: for each subnetwork the number of occurrences of each word w in the corresponding subnetwork i is accumulated in an array $N(w, i)$, and normalised by dividing $N(w, i)$ by the number of combined systems (N_s) to scale the frequency of occurrence to the unity. Moreover, depending on the voting scheme, the confidence scores for word w in the subnetwork i are measured and normalised in an array $C(w, i)$. The confidence score of NULL transition arcs can be defined by the Conf(@) parameter.

The balance between using word frequency and confidence scores can be adjusted by means of a parameter α:

$$\mathrm{Score}(w, i) = \alpha\left(\frac{N(w, i)}{N_s}\right) + (1 - \alpha)C(w, i) \tag{4}$$

The voting search module offers the following three different voting schemes:

1. **Frequency of occurrence.** In the voting by frequency of occurrence scheme all confidence scoring information is ignored, i.e. the α parameter is set to 1.
2. **Frequency of occurrence and average word confidence.** In this voting method, the confidence score of each word w in the array $C(w, i)$ is set to the average value of the appearance of this word w in the subnetwork i. Both parameters α and Conf(@) must be trained a priori.
3. **Frequency of occurrence and maximum confidence.** In the last voting scheme, the confidence score of each word w in the array $C(w, i)$ is set to the maximum value of the appearance of this word w in the subnetwork i. In this case, both parameters α and Conf(@) must be also trained a priori.

3.2.2. N-best ROVER

The combination of multiple hypotheses can produce an output more accurate than combining only the 1-best hypothesis. This is the idea behind the N-best ROVER method (Stolcke et al., 2000), which uses n-best outputs to perform the combination.

This method works in three steps. In a first step, the n-best h hypotheses from the decoding of a feature vector sequence x by using different systems S_i are aligned like in the ROVER method. Then, in a second step the normalised and weighted log-linear word posteriors are estimated for each system. In the last step, the combined word posterior is computed as a linear combination.

The word posteriors for each word w and system i are computed for each subnetwork j by log-linear score weighting, followed by a normalisation over all hypotheses.

$$\mathrm{P}_i(w \mid x) = \frac{\sum\limits_{h:w\in h} \exp\left(\sum\limits_{j} \lambda_{ij} s_{ij}(h \mid x)\right)}{\sum\limits_{\forall h} \exp\left(\sum\limits_{j} \lambda_{ij} s_{ij}(h \mid x)\right)} \tag{5}$$

where $s_{ij}(h \mid x)$ is the log-score, and λ_{ij} are the log-score combination weights for the subnetwork j of the hypothesis h of the system i. Then, the combined posterior can be computed as a linear combination:

$$\mathrm{P}(w \mid x) = \sum_{i} \mu_i \mathrm{P}_i(w \mid x) \tag{6}$$

where μ_i represents the system weight.

Finally, the combined hypothesis is formed by the concatenation of the most probable word hypotheses at each position in the alignment. Therefore, like ROVER, this method presents the following constraint: the result of the combination is composed of a single hypothesis.

3.2.3. Lattices Rescoring

Combining multiple lattices on a new lattice not only may improve the most likely hypothesis, but also this new lattice may contain better hypotheses than the most likely. The N-best List and Lattices Rescoring method (Ostendorf et al., 1991; Stolcke et al., 1997) optimises the word-level recognition scores and constructs a word lattice from all information contained in the lattices to combine.

This algorithm has two components. In the first one, the scores of the hypotheses contained in the lattices to combine are weighted by using a parameter, and then all these hypotheses are aligned and merged in one n-best list. In the second one, the optimisation of the word-level recognition scores is made by means of the substitution of the normalisation term P(x) of Equation (1) by a finite sum over the set W of all the hypotheses in the joint n-best list:

$$\mathrm{P}(x) = \sum_{w\in W} \mathrm{P}(w \mid x) \tag{7}$$

Finally, a new combined lattice is build from the rescored n-best hypotheses.

3.2.4. Confusion Network Combination

Another way to obtain a combined lattice is the use of Confusion Network Combination (CNC) methods. The bimodal CNC method presented in (Granell and Martínez-Hinarejos, 2015a) is a special case of CNC where the hypotheses of one modality are used to minimise the word error present in the other modality hypotheses. In a first step, both CN are aligned by similarity based on a gram matching error. The gram matching between the words of both modalities (w_A and w_B) is assessed by using the quadratic mean of the Character Error Rate (CER) and the Phoneme Error Rate (PER) between those words:

$$\mathrm{E}(w_A, w_B) = \sqrt{\frac{\mathrm{CER}(w_A, w_B)^2 + \mathrm{PER}(w_A, w_B)^2}{2}} \tag{8}$$

Where CER and PER are the Levensthein distance between the words of both modalities, CER at character level, and PER at phoneme level by using the phonetic transcriptions of the recognised words, and E represents the gram matching error.

In a second step a new CN is composed on the basis of the Bayes theorem, assuming a strong independence between both modalities, by using three editing actions: combination, insertion, and deletion of subnetworks. Given two subnetworks, SN_A and SN_B, the word posterior probabilities of the combined subnetwork SN_C are obtained by applying a normalisation on the logarithmic interpolation of the smoothed word posterior probabilities of both subnetworks:

$$\mathrm{P}(w \mid SN_C) = \mathrm{P}(w \mid SN_A)^{\alpha}\mathrm{P}(w \mid SN_B)^{1-\alpha} \quad (9)$$

For insertion and deletion, the subnetwork to insert or to delete is combined with a subnetwork with an only *DELETE* arc with probability 1.0.

4. Experimental Framework

In this section, the used data, and the system setup (features, models, and evaluation metrics) are presented.

4.1. Corpora

Two handwritten document datasets and three spoken corpora were used in the experiments. All these corpora are described in the following parts.

4.1.1. Handwritten text: Cristo Salvador

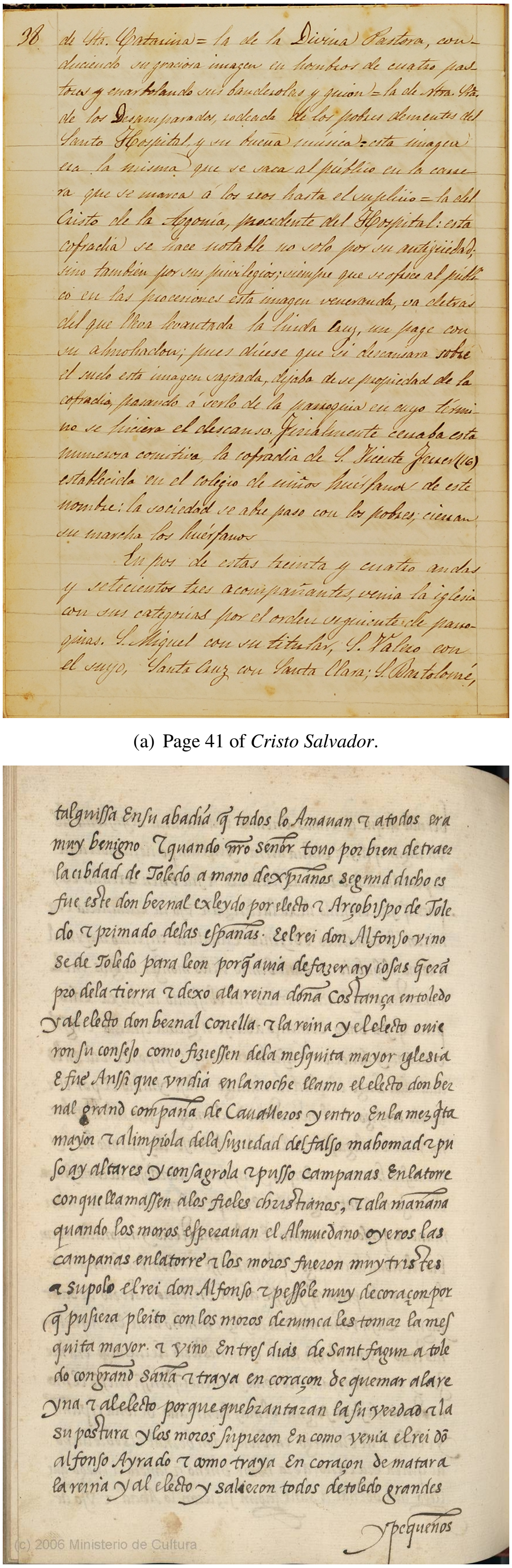
The *Cristo Salvador* corpus was employed previously in different works (Alabau et al., 2011, 2014; Granell and Martínez-Hinarejos, 2015a) related to multimodal combination. This corpus is a handwritten book of the XIX century provided by *Biblioteca Valenciana Digital* (BiValDi). It is a single writer book with different image features that have some problems, such as smear, background variations, differences in bright, and bleed-through (ink that trespasses to the other surface of the sheet). It is composed of 53 pages (page 41 is presented in Figure 3(a)). This corpus is available with the pages divided into lines (such as shown in Figure 4).

This corpus presents a total number of 1,172 lines, with a vocabulary of 3,287 different words. For training the optical models for this HTR corpus, a partition with the first 32 pages (675 lines) was used. The obtained optical models modelled the set of 78 symbols present in this corpus, taking into account lowercase and uppercase letters, numbers, punctuation marks, special symbols, and blank spaces.

Test data was composed of the lines of page 41 (24 lines, 222 words), that was selected for being, according to preliminary error recognition results, a representative page of the whole test set (the remaining 21 pages, 497 lines). The multimodal test set is composed of a single page because the speech acquisition would not be feasible for the whole standard test set.

4.1.2. Handwritten text: Rodrigo

The *Rodrigo* corpus (Serrano et al., 2010) is composed of a set of 853 pages written by a single writer in 1545, entitled "Historia de España del arçobispo Don Rodrigo". The topic of



38 de Sta. Catalina = la de la Divina Pastora, con-
duciendo su graciosa imagen en hombros de cuatro pas-
tores y enarbolando sus banderolas y guion = la de Ntra. Sra.
de los Desamparados, rodeada de los pobres dementes del
Santo Hospital, y su buena música: esta imagen
era la misma que se saca al público en la carre-
ra que se marca á los reos hasta el suplicio = la del
Cristo de la Agonía, procedente del Hospital: esta
cofradía se hace notable no solo por su antigüedad;
sino tambien por sus privilegios; siempre que se ofrece al públi-
co en las procesiones esta imagen veneranda, va detras
del que lleva levantada la lucida cruz, un page con
su almohadon; pues dícese que si descansara sobre
el suelo esta imagen sagrada, dejaba de ser propiedad de la
cofradía, pasando á serlo de la parroquia en cuyo térmi-
no se hiciera el descanso. Finalmente cerraba esta
numerosa comitiva, la cofradía de S. Vicente Ferrer (16)
establecida en el colegio de niños huérfanos de este
nombre: la sociedad se abre paso con los pobres; cierran
su marcha los huérfanos

En pos de estas treinta y cuatro andas
y setecientos tres acompañantes, venia la iglesia
con sus categorias por el orden siguiente de parro-
quias. S. Miguel con su titular, S. Valero con
el suyo, Santa Cruz con Santa Clara; S. Bartolomé,

(a) Page 41 of *Cristo Salvador*.

talguiſſa Enſu abadía q todos lo Amauan ⁊ atodos era
muy benigno ⁊ quando nro señor touo por bien detraer
la cibdad de Toledo a mano dexpianos segund dicho es
fue este don bernal exleydo por electo ⁊ Arçobiſpo de Tole
do ⁊ primado delas españas · Eelrei don Alfonſo vino
se de Toledo para leon porq auia defazer ay coſas q erā
pro dela tierra ⁊ dexo ala reina doña Costança en toledo
y al electo don bernal conella · ⁊ la reina y el electo ouie
ron ſu conſejo como fizieſſen dela meſquita mayor yglesia
E fue Anſſi que vn dia enlanoche llamo el electo don ber
nal grand compaña de Caualleros y entro Enla mezq̃ta
mayor ⁊ alimpiola dela ſuziedad del falſo mahomad ⁊ pu
ſo ay altares y conſagrola ⁊ puſſo campanas Enlatorre
con que llamaſſen alos fieles christianos, ⁊ ala mañana
quando los moros eſperauan el Almuedano oyeron las
campanas enlatorre ⁊ los moros fueron muy tristes
⁊ supolo elrei don Alfonſo ⁊ peſſole muy de coraçon por
q puſiera pleito con los moros de nunca les tomar la meſ
quita mayor · ⁊ vino En treſ dias de Sant fagun a tole
do con grand saña ⁊ traya en coraçon de quemar ala re
yna ⁊ al electo porque quebrantaran la ſu verdad ⁊ la
su postura y los moros ſupieron En como venia elrei dō
alfonſo Ayrado ⁊ como traya En coraçon de matar a
la reina y al electo y salieron todos de toledo grandes
y pequeños

(b) Page 515 of *Rodrigo*.

Fig. 3. A page example for each handwritten text corpus.

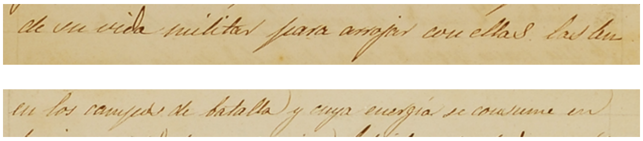


Fig. 4. Examples of *Cristo Salvador* handwritten lines.

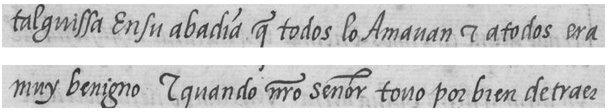


Fig. 5. Examples of *Rodrigo* handwritten lines.

the book is historical chronicles of Spain. Most pages (as page 515 shown in Figure 3(b)) form a single block with well separated lines (usually 25 lines per page), written in calligraphical text. The corpus is available with the separated lines, which are the source for feature extraction (Figure 5 shows an example of separated lines). The corpus has a total of 20,356 lines in PNG format and a vocabulary size of about 11,000 words. For training the optical models, a standard partition with a total number of 5000 lines (about 205 pages) was used.

The test set for this HTR corpus was composed of two pages that were not included in the training part (pages 515 and 579) and that were representative of the average error of the standard test set (of about 5000 lines). These two pages contain 50 lines and 514 words. As happened with *Cristo Salvador*, these two pages where selected to allow a reasonable speech acquisition for the multimodal experiments. The set of 106 symbols present in this corpus gets modelled by the corresponding optical models, which take into account lowercase and uppercase letters, numbers, punctuation marks, special symbols, and blank spaces.

4.1.3. *Speech: Albayzin, Cristo Salvador and Rodrigo*

For the training of the ASR acoustical models we used a partition of the Spanish phonetic corpus Albayzin (Moreno et al., 1993). This corpus consists of a set of three sub-corpus recorded by 304 speakers using a sampling rate of 16 KHz and a 16 bit quantisation. The training partition used in this work includes a set of 4800 phonetically balanced utterances, specifically, 200 utterances read by four speakers and 25 utterances read by 160 speakers, with a total length of about 4 hours. Acoustical models cover a total of 25 phones (23 monophones, short silence, and long silence), and they were estimated from this corpus.

Test data for ASR was the product of the acquisition of the dictation of the contents of the lines of the test pages of each handwritten corpus using a sample rate of 16 KHz and an encoding of 16 bits (to match the conditions of Albayzin data). In the case of *Cristo Salvador*, the ASR test data was composed of the acquisition of the lines of the page 41 by five different native Spanish speakers (i.e., a total set of 120 utterances, with a total length of about 9 minutes), while in the case of *Rodrigo*, seven different native Spanish speakers read the 50 handwritten test lines (those of pages 515 and 579), giving a total set of 350 utterances (about 15 minutes).

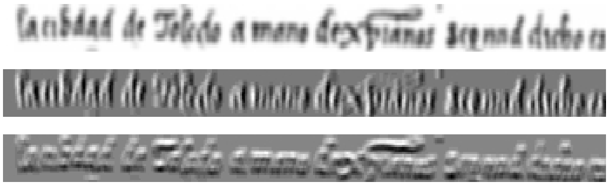


Fig. 6. Example of feature vectors sequence.

4.2. *System setup*

In this work we employed and compared systems based on HMM and on deep learning models. The HTR and ASR recognition systems based on HMM were implemented by using the iATROS recogniser (Luján-Mares et al., 2008), and the HTR and ASR recognition systems based on deep learning were implemented by using the Laia (Puigcerver et al., 2016) and EESEN (Miao et al., 2015) recognisers. The SRILM toolkit (Stolcke, 2002) was used to obtain CN from the WG recognition outputs.

4.2.1. *Features*

Handwritten text features are computed in several steps from line images. These steps are different depending on the final models used. Common steps are slant correction by using the maximum variance method (Pastor et al., 2004) and a size normalisation. In the case of the deep learning models, the input is the line image after performing these common steps and the process described in (Villegas et al., 2015).

However, in the case of HMM a more complex sequence of steps is used. First, a bright normalisation is performed. After that, a median filter of size 3 × 3 pixels is applied to whole image. Next, the common steps (slant correction and size normalisation) are aplied. Finally, each preprocessed line image is represented as a sequence of feature vectors. To do this, the text line image is divided into squared cells. From each cell, three features are calculated: normalised grey level, horizontal grey level derivative and vertical grey level derivative. Columns of cells (frames) are processed from left to right and a feature vector is constructed for each frame by stacking the three features computed in its constituent cells. The way these three features are determined is described in (Toselli et al., 2004). In this work, final feature vectors for HMM are of 60 dimensions. In Figure 6 we can see an example of the feature vectors sequence obtained from the image in the CATTI example (Figure 2).

With respect to speech features, Mel-Frequency Cepstral Coefficients (MFCC) are extracted from the audio files. The Fourier transform is calculated every 10 ms over a window of 25 ms of a pre-emphasised signal. Next, 23 equidistant Mel scale triangular filters are applied and the filters outputs are logarithmised. Finally, to obtain the MFCC a discrete cosine transformation is applied. In this work, the first 12 MFCC and log frame energy with first and second order derivatives are used, resulting in a 39 dimensional feature vector (Rabiner and Juang, 1993).

Table 1. Architecture of the optical models based on deep neural networks.

Parameters	*Cristo Salvador*	*Rodrigo*
CNN Layers	3	5
Filters	$\{16,32,48\}$	$\{16,32,48,64,80\}$
Kernel size	3×3	3×3
MaxPool size	2×2	2×2
Dropout	$\{0,0.2,0.2\}$	$\{0,0,0.2,0.2,0.2\}$
RNN Layers	3	5
BLSTM Units	256	256

4.2.2. Models

Optical and acoustical HMM models were trained by using HTK (Young et al., 2006). On the one hand, symbols on the optical models are modelled by a continuous density left-to-right HMM with 14 and 4 states for *Cristo Salvador* and *Rodrigo*, respectively, and 32 gaussians per state. On the other hand, phonemes on the acoustical model are modelled as a left-to-right HMM with 3 states and 64 gaussians per state.

Optical models based on deep learning were trained by using Laia (Puigcerver et al., 2016). Those optical models are Convolutional Recurrent Neural Networks (CRNN) (Choi et al., 2016), which consist of a convolutional (CNN) and a recurrent (RNN) blocks with the architecture detailed in Table 1 for each corpus. The convolutional blocks, are composed of convolutional layers with filters composed by different features maps. Each convolutional layer has kernel sizes of 3×3 pixels, horizontal and vertical strides of 1 pixel, LeakyReLU as activation function, and a maximum pooling layer with non-overlapping kernels of 2×2 pixels only at the output of the first two layers for *Cristo Salvador* and at the output of the first three layers for *Rodrigo*. Then, the recurrent blocks are composed of different recurrent layers composed of 256 Bidirectional Long-Short Term Memory (BLSTM) units. Finally, a linear fully-connected output layer is used after the recurrent block. The first part of the architechture is very similar to that employed in OCR and scene text detection and recognition, where CNN blocks are employed for single character recognition (Wang et al., 2012; Jaderberg et al., 2014b). The second part of the architechture (the recurrent layers) allow to model the connection of the cursive text.

Acoustical models based on deep learning were trained by using EESEN (Miao et al., 2015). This acoustical deep model is a Recurrent Neural Network (RNN) composed of 351 inputs for 9 neighbouring frames of cepstral features, 6 hidden layers with 250 BLSTM units, and an output layer with a softmax function (Graves et al., 2013).

The lexicon models for both systems are in HTK lexicon format, where each word is modelled as a concatenation of symbols for HTR or phonemes for ASR.

The language models (LM) were estimated directly from the transcriptions of the pages included on the HTR training sets (32 pages for *Cristo Salvador*, and about 205 pages for *Rodrigo*) by using the SRILM *ngram-count* tool (Stolcke, 2002). The models were 2-gram with Kneser-Ney back-off smoothing (Kneser and Ney, 1995). The *Cristo Salvador* language model was interpolated with the whole lexicon in order to avoid out-of-vocabulary words, and it presents a perplexity of 742.8 for the test data. In contrast, the *Rodrigo* language model presents for the test a 6.2% of out-of-vocabulary words and a perplexity of 298.4. Although language models could be enriched with external sources, the antiquity and the topic of the books makes difficult to find representative texts to enhance the language models.

4.2.3. Evaluation Metrics

Different evaluation measures have been adopted. On the one hand, the quality of the transcription without any system-user interactivity is given by the well known word error rate (WER), which is a good estimation of the user post-edition effort. It is defined as the minimum number of words to be substituted, deleted or inserted to convert the hypothesis into the reference, divided by the total number of reference words. In addition, the oracle WER represents the best WER that can be obtained from a lattice.

On the other hand, the CATTI performance is given by the word stroke ratio (WSR), which can be also computed using the reference transcription. After each CATTI hypothesis, the longest common prefix between the hypothesis and the reference is obtained and the first mismatching word from the hypothesis is replaced by the corresponding reference word. This process is iterated until a full match is achieved. Therefore, the WSR can be defined as the number of user interactions that are necessary to produce correct transcriptions using the CATTI system, divided by the total number of reference words. This definition makes WER and WSR comparable. The relative difference between them gives us a good estimation of the reduction in human effort (EFR) that can be achieved by using CATTI with respect to using a conventional HTR system followed by human post-editing.

For both measures WER and WSR, confidence intervals of 95% were calculated by using the bootstrapping method with 10,000 repetitions (Bisani and Ney, 2004). In order to confirm the statistical significance, p-values with a threshold of significance of $\alpha = 0.05$ were calculated through the Welch t-test (Welch, 1947) by using the statistical computing tool R (R Core Team, 2017).

5. Experimental Results

Multimodal combination allows to enrich the CATTI hypotheses from different sources of information (in this case HTR and ASR). Several experiments were performed in order to test our multimodal proposal by using two different handwritten text datasets (*Cristo Salvador* and *Rodrigo*). Two different decoding approaches were tested, traditional HTR and ASR based on Hidden Markov Models (HMM) and the state-of-the-art HTR and ASR based on deep learning. Moreover, the performance of this multimodal proposal was tested by using the four different combination techniques described in Section 3.

For both HTR corpora and decoding approaches, the unimodal post-edition baseline values were obtained. Next, the classical unimodal CATTI was tested. Then, both modalities were combined by using the different combination techniques. Finally, the new multimodal CATTI proposal was tested.

In order to optimise the experimental results, the values of the main decoding parameters (beam, word insertion penalty, ...) were tuned. In the CATTI experiments, the limit of mouse actions was set to 3. The different combination parameters were: frequency of occurrence voting scheme for ROVER (since the multimodal combination is performed without training), uniform weights ($\lambda = 1$ and $\mu = 0.5$) for N-best ROVER, and combination weight of 0.5 for the Lattices Rescoring and CN Combination techniques.

5.1. HMM Based Decoding for Cristo Salvador

Table 2 presents the obtained general HMM based results for the *Cristo Salvador* corpus. As can be observed in the unimodal post-edition results, speech recognition does not seem to be a good substitute for handwriting recognition in this task. However, the ASR oracle WER value is similar to the HTR oracle WER.

Regarding the unimodal CATTI results presented in the top-right part of the table, the estimated interactive human effort (WSR) required for obtaining the perfect transcription from the HTR decoding represents 8.2% of relative effort reduction (EFR) over the HTR baseline WER (p-value = .541). However, no effort reduction can be considered when only ASR is used at the input of the CATTI system.

As expected, in the multimodal experiments the use of all hypotheses in the combination (used by Lattices Rescoring and CN Combination) allows to obtain better post-edition results than using only a single hypothesis (which is the limitation of ROVER and N-best ROVER). The best result was obtained by using the CN Combination method with a 29.3% ± 2.5 of WER, which represents a relative improvement over the HTR baseline WER (32.9% ± 6.8) of 10.9% (p-value = .296). The best oracle WER (10.3%) was obtained by using the Lattices Rescoring method. Results show that WER improvements are not significant with respect to HTR baseline WER, but the oracle WER values are substantially lower in the case of lattice combination methods. Therefore, an outstanding effect of multimodal lattices combination in interactive transcription systems can be expected, since this low oracle WER is related to the amount and quality of the alternatives offered by the combination technique (the lower the oracle WER, the more and better alternatives).

Regarding the obtained results in the multimodal CATTI approach, the use of the ROVER and N-best ROVER combination methods produce worse results when comparing with the unimodal HTR CATTI baseline WSR (30.2% ± 6.4), although differences are no statistically significant (p-value > .010). In contrast, the values obtained by the CN Combination and Lattices Rescoring methods not only represent improvements, but these improvements are also statistically significant (p-value < .001). Concretely, the overall best result (13.7% ± 2.0) was achieved by using the Lattices Rescoring method and it represents a relative improvement of 54.6% over the unimodal HTR CATTI WSR, and an EFR of 58.4% over the unimodal HTR baseline WER.

5.2. Deep Learning Based Decoding for Cristo Salvador

The obtained results for the deep learning based decoding experiments for the *Cristo Salvador* corpus are presented in Table 3. In the unimodal post-edition results, the deep learning based decoding offers statistically significant (p-value < .001) better transcriptions that the HMM based decoding, for both modalities, although the results of ASR are still quite poor.

In the unimodal CATTI experiments, only in the HTR case some effort reduction can be considered. Concretely, it presents a WSR equal to 4.1%, which represents 53.9% of relative effort reduction over the HTR baseline (WER equal to 8.9%, p-value = .051).

In the multimodal experiments, the best result was obtained by using the CN Combination method with a 8.4% ± 1.6 of WER, which represents a relative improvement over the HTR baseline WER (8.9% ± 4.2) of 5.6% (p-value = .864). The best oracle WER (0.2%) was obtained by using the Lattices Rescoring method. Although these WER improvements are not statistically significant, the oracle WER values obtained in the case of lattice combination methods are exceptionally low.

In the multimodal CATTI approach, the use of the ROVER and N-best ROVER combination methods produce worse results when comparing with the unimodal HTR CATTI baseline WSR (4.1% ± 2.9). On the other side, the use of CN Combination and Lattices Rescoring methods improves the performance. The overall best result (1.8% ± 0.6) was achieved by using the Lattices Rescoring method and it represents a relative improvement of 56.1% over the unimodal HTR CATTI WSR (p-value = .091), and an EFR of 79.8% over the unimodal HTR baseline (WER value equal to 8.9%, p-value < .001).

5.3. HMM Based Decoding for Rodrigo

The same procedure was followed with this corpus. In Table 4 the obtained general HMM decoding results are shown. With the post-edition experiments, we confirmed that speech recognition is not a good substitute for historical handwriting recognition. However, as happened with the HMM decoding in *Cristo Salvador*, for this corpus both modalities also present similar oracle WER values.

The obtained WSR value (36.2% ± 3.6) in the unimodal HTR CATTI experiment represents a relative effort reduction of 7.9% over the *Rodrigo* unimodal HTR baseline (WER equal to 39.3% ± 4.1, p-value = .304). Nevertheless, in the case of unimodal ASR neither can any effort reduction be considered.

Regarding the multimodal results, in post-edition all techniques present similar performances, except ROVER, that present a statistically significant worse result (p-value = .016). However, in the CATTI experiments Lattices Rescoring and CN Combination methods presented significantly better results than ROVER and N-best ROVER (p-value < .001), which is in consonance with their oracle WER results.

In the post-edition experiments, the best result (35.9% ± 1.6) was obtained by using the CN Combination method, and it represents 8.7% of relative improvement over the HTR baseline (WER equal to 39.3% ± 4.1, p-value = .138). Meanwhile, the Lattices Rescoring method allowed to obtain the best oracle WER (10.6%).

Table 2. ***Cristo Salvador*** **HMM Based Experimental Results. The relative human effort reduction (EFR) represents the relative difference between the obtained CATTI WSR over the unimodal HTR post-edition WER value.**

Experiment	Post-edition		CATTI	
	WER	Oracle WER	WSR	EFR
Unimodal HTR	32.9% ± 6.8	27.5%	30.2% ± 6.4	8.2%
Unimodal ASR	43.3% ± 3.4	27.4%	35.1% ± 3.5	−6.7%
Multimodal (ROVER)	32.7% ± 2.9	32.7%	32.8% ± 2.6	0.3%
Multimodal (N-best ROVER)	33.3% ± 2.9	33.3%	35.9% ± 2.6	−9.1%
Multimodal (Lattices Rescoring)	31.3% ± 2.6	**10.3%**	**13.7% ± 2.0**	**58.4%**
Multimodal (CN Combination)	**29.3% ± 2.5**	13.4%	14.1% ± 2.2	57.1%

Table 3. ***Cristo Salvador*** **Deep Learning Based Experimental Results. The relative human effort reduction (EFR) represents the relative difference between the obtained CATTI WSR over the unimodal HTR post-edition WER value.**

Experiment	Post-edition		CATTI	
	WER	Oracle WER	WSR	EFR
Unimodal HTR	8.9% ± 4.2	1.8.%	4.1% ± 2.9	53.9%
Unimodal ASR	31.4% ± 3.4	8.5%	10.4% ± 2.0	−16.9%
Multimodal (ROVER)	14.0% ± 2.5	14.0%	7.7% ± 2.2	13.5%
Multimodal (N-best ROVER)	8.8% ± 1.8	8.8%	9.0% ± 2.0	−1, 1%
Multimodal (Lattices Rescoring)	8.6% ± 1.7	**0.2%**	**1.8% ± 0.6**	**79.8%**
Multimodal (CN Combination)	**8.4% ± 1.6**	0.3%	2.7% ± 1.1	69.7%

Table 4. ***Rodrigo*** **HMM Based Experimental Results. The relative human effort reduction (EFR) represents the relative difference between the obtained CATTI WSR over the unimodal HTR post-edition WER value.**

Experiment	Post-edition		CATTI	
	WER	Oracle WER	WSR	EFR
Unimodal HTR	39.3% ± 4.1	28.0%	36.2% ± 3.6	7.9%
Unimodal ASR	62.9% ± 2.2	29.5%	47.2% ± 2.3	−20.1%
Multimodal (ROVER)	44.9% ± 1.8	44.9%	44.8% ± 1.7	−14.0%
Multimodal (N-best ROVER)	38.4% ± 1.8	38.4%	41.0% ± 1.8	−4.3%
Multimodal (Lattices Rescoring)	37.2% ± 1.7	**10.6%**	**25.2% ± 1.6**	**35.9%**
Multimodal (CN Combination)	**35.9% ± 1.6**	14.8%	27.0% ± 1.8	31.3%

On the other hand, the use of the ROVER and N-best ROVER combination methods on the multimodal CATTI does not improve the unimodal HTR baseline WSR (36.2% ± 3.6). However, the CN Combination and Lattices Rescoring combination methods allow to obtain statistically significant improvements with an EFR higher than 30% over the HTR baseline WER (p-value < .001). The Lattices Rescoring combination method allowed to obtain the overall best WSR result, specifically 25.2% ± 1.6 of WSR, which represents a relative improvement of 30.4% over the unimodal HTR baseline WSR (p-value < .001).

5.4. *Deep Learning Based Decoding for* Rodrigo

Table 5 presents the obtained results for the deep learning based decoding experiments for the *Rodrigo* corpus. As for the *Cristo Salvador* corpus, the deep learning based decoding offers statistically significant (p-value < .001) better transcriptions that the HMM based decoding, for both modalities. However, the results of ASR are still quite poor due to the difficulty of the task.

In the unimodal CATTI experiments, only in the HTR case some effort reduction can be considered. Concretely, it presents a WSR equal to 8.4%, which represents 29.9% of relative effort reduction over the HTR baseline (WER equal to 12.0%, p-value = .063).

Unlike previous multimodal experiments, in this case, the best results were obtained by using the Lattices Rescoring method with a 11.9% ± 1.2 of WER, and 6.6% of oracle WER.

In the multimodal CATTI approach, only the use of the Lattices Rescoring method produce better results when comparing with the unimodal HTR CATTI baseline WSR (8.4% ± 2.3). Specifically, 7.6% ± 0.9 of WSR, which represents a relative improvement of 9.5% over the unimodal HTR CATTI WSR (p-value = .550), and an EFR of 36.6% over the unimodal HTR baseline (WER equal to 12.0%, p-value < .005).

6. Conclusions

In this paper, we have proposed the use of multimodal combination techniques for improving the CATTI system presented in previous works.

Table 5. *Rodrigo* Deep Learning Based Experimental Results. The relative human effort reduction (EFR) represents the relative difference between the obtained CATTI WSR over the unimodal HTR post-edition WER value.

Experiment	Post-edition		CATTI	
	WER	Oracle WER	WSR	EFR
Unimodal HTR	12.0% ± 3.1	8.4%	8.4% ± 2.3	29.9%
Unimodal ASR	50.4% ± 2.3	22.3%	19.5% ± 1.7	−62.6%
Multimodal (ROVER)	23.2% ± 1.8	23.2%	15.8% ± 1.7	−31.8%
Multimodal (N-best ROVER)	13.0% ± 1.3	13.0%	9.7% ± 0.9	19.1%
Multimodal (Lattices Rescoring)	**11.9% ± 1.2**	**6.6%**	**7.6% ± 0.9**	**36.6%**
Multimodal (CN Combination)	13.4% ± 1.5	8.0%	8.6% ± 0.9	28.2%

By means of multimodal combination techniques, we have confirmed the benefits of using speech as an additional source of information for the assisted transcription of historical manuscripts.

The use of lattice combination techniques permits to obtain transcription outputs with a reduced error. This error reduction is due to the fact that the combination may produce new bigrams that increase the search alternatives, and that the adjustment of the word posterior probabilities can increase the probabilities of the correct words. The main advantage of the presented approach is that the error reduction produced by lattice combination techniques allows to reduce significantly the human effort when using an assistive transcription system.

The obtained results show that there is still room for improvement. We propose for future improvement the use of sentences in the handwritten text corpus instead of lines, in order to make multimodality more natural. Moreover, our future work aims at the improvement of the user interaction with more ergonomic feedback modalities, such as on-line handwritten text recognition.

Acknowledgments

Work partially supported by projects READ - 674943 (European Union's H2020), SmartWays - RTC-2014-1466-4 (MINECO), and CoMUN-HaT - TIN2015-70924-C2-1-R (MINECO / FEDER), and by Generalitat Valenciana (GVA) under reference PROMETEOII/2014/030.

References

Alabau, V., Martínez-Hinarejos, C.D., Romero, V., Lagarda, A.L., 2014. An iterative multimodal framework for the transcription of handwritten historical documents. Pattern Recognition Letters 35, 195–203. Frontiers in Handwriting Processing.

Alabau, V., Romero, V., Lagarda, A.L., Martínez-Hinarejos, C.D., 2011. A Multimodal Approach to Dictation of Handwritten Historical Documents., in: Proc. 12th Interspeech, pp. 2245–2248.

Barrachina, S., Bender, O., Casacuberta, F., Civera, J., Cubel, E., Khadivi, S., Lagarda, A., Ney, H., Tomás, J., Vidal, E., et al., 2009. Statistical approaches to computer-assisted translation. Computational Linguistics 35, 3–28.

Bisani, M., Ney, H., 2004. Bootstrap estimates for confidence intervals in ASR performance evaluation, in: Proc. of ICASSP, pp. 409–412.

Bissacco, A., Cummins, M., Netzer, Y., Neven, H., 2013. Photoocr: Reading text in uncontrolled conditions, in: 2013 IEEE International Conference on Computer Vision (ICCV), pp. 785–792. URL: doi.ieeecomputersociety.org/10.1109/ICCV.2013.102, doi:10.1109/ICCV.2013.102.

Bloomberg, D.S., Kopec, G.E., Dasari, L., 1995. Measuring document image skew and orientation. SPIE 2422, 302–316.

Choi, K., Fazekas, G., Sandler, M.B., Cho, K., 2016. Convolutional recurrent neural networks for music classification. CoRR abs/1609.04243. URL: http://arxiv.org/abs/1609.04243.

Di Fabbrizio, G., Okken, T., Wilpon, J.G., 2009. A Speech Mashup Framework for Multimodal Mobile Services, in: Proc. of ICMI-MLMI '09, pp. 71–78. URL: http://doi.acm.org/10.1145/1647314.1647329, doi:10.1145/1647314.1647329.

Doi, K., 2007. Computer-aided diagnosis in medical imaging: Historical review, current status and future potential. Computerized Medical Imaging and Graphics 31, 198 – 211. doi:http://dx.doi.org/10.1016/j.compmedimag.2007.02.002.

Dupont, S., Luettin, J., 2000. Audio-visual speech modeling for continuous speech recognition. IEEE transactions on multimedia 2, 141–151.

Evermann, G., Woodland, P., 2000. Posterior probability decoding, confidence estimation and system combination, in: Proc. of Speech Transcription Workshop.

Fischer, A., Wuthrich, M., Liwicki, M., Frinken, V., Bunke, H., Viehhauser, G., Stolz, M., 2009. Automatic Transcription of Handwritten Medieval Documents, in: Proc. of VSMM '09, pp. 137–142. doi:10.1109/VSMM.2009.26.

Fiscus, J.G., 1997. A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER), in: Proc. of ASRU 1997, pp. 347–354.

Gordo, A., Llorens, D., Marzal, A., Prat, F., Vilar, J.M., 2008. State: A Multimodal Assisted Text-Transcription System for Ancient Documents, in: Proc. of the 8th IAPR-DAS, pp. 135–142. URL: http://dx.doi.org/10.1109/DAS.2008.28, doi:10.1109/DAS.2008.28.

Granell, E., Martínez-Hinarejos, C.D., 2015a. Combining handwriting and speech recognition for transcribing historical handwritten documents, in: Proc. of the 13th ICDAR, pp. 126–130.

Granell, E., Martínez-Hinarejos, C.D., 2015b. Multimodal Output Combination for Transcribing Historical Handwritten Documents, in: Proc. of the 16th CAIP, pp. 246–260.

Granell, E., Romero, V., Marínez-Hinarejos, C.D., 2016. An Interactive Approach with *Off-line* and *On-line* Handwritten Text Recognition Combination for Transcribing Historical Documents, in: Proc. of the 12th IAPR-DAS, pp. 269–274.

Graves, A., Mohamed, A.r., Hinton, G., 2013. Speech recognition with deep recurrent neural networks, in: Acoustics, speech and signal processing (icassp), 2013 ieee international conference on, IEEE. pp. 6645–6649.

Grüning, T., Leifert, G., Strauß, T., Labahn, R., 2018. A two-stage method for text line detection in historical documents. CoRR abs/1802.03345. URL: http://arxiv.org/abs/1802.03345.

Hernando, J., Ayarte, J., Monte, E., 1995. Optimization of speech parameter weighting for cdhmm word recognition, in: Proc. 4th EUROSPEECH, pp. 105–108.

Jaderberg, M., Simonyan, K., Vedaldi, A., Zisserman, A., 2014a. Synthetic data and artificial neural networks for natural scene text recognition. CoRR abs/1406.2227. URL: http://arxiv.org/abs/1406.2227.

Jaderberg, M., Vedaldi, A., Zisserman, A., 2014b. Deep features for text spotting, in: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (Eds.), Computer Vision – ECCV 2014, Springer International Publishing. pp. 512–528.

Jelinek, F., 1998. Statistical Methods for Speech Recognition. MIT Press.

Jurafsky, D., Martin, J.H., 2009. Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics. Prentice Hall.

Kneser, R., Ney, H., 1995. Improved backing-off for m-gram language modeling, in: Proc. of ICASSP, pp. 181–184.

Li, X., 2005. Combination and generation of parallel feature streams for improved speech recognition. Ph.D. thesis. Carnegie Mellon University Pittsburgh, PA.

Luján-Mares, M., Tamarit, V., Alabau, V., Martínez-Hinarejos, C.D., Pastor, M., Sanchis, A., Toselli, A.H., 2008. iATROS: A speech and handwritting recognition system, in: V Jornadas en Tecnologías del Habla, pp. 75–78.

Machover, C., 1995. The CAD/CAM Handbook. McGraw-Hill.

Malit, R.F., 2009. Computer assisted driving of vehicles. URL: http://www.google.tl/patents/US7513508. uS Patent 7,513,508.

Miao, Y., Gowayyed, M., Metze, F., 2015. EESEN: End-to-end speech recognition using deep RNN models and WFST-based decoding, in: Proc. of ASRU 2015, IEEE. pp. 167–174.

Mihalcea, R., 2012. Multimodal Sentiment Analysis, in: Proc. of WASSA '12, pp. 1–1. URL: http://dl.acm.org/citation.cfm?id=2392963.2392965.

Moreno, A., Poch, D., Bonafonte, A., Lleida, E., Llisterri, J., Mariño, J.B., Nadeu, C., 1993. Albayzin speech database: design of the phonetic corpus, in: Proc. of EuroSpeech, pp. 175–178.

Ostendorf, M., Kannan, A., Austin, S., Kimball, O., Schwartz, R.M., Rohlicek, J.R., 1991. Integration of diverse recognition methodologies through reevaluation of n-best sentence hypotheses., in: HLT, pp. 83–87.

Pastor, M., Toselli, A.H., Vidal, E., 2004. Projection profile based algorithm for slant removal, in: Proc. of ICIAR'04, pp. 183–190.

Potamianos, G., Neti, C., 2001. Automatic Speechreading Of Impaired Speech, in: Proc. of AVSP, pp. 177–182.

Potamianos, G., Neti, C., Gravier, G., Garg, A., Senior, A.W., 2003. Recent advances in the automatic recognition of audiovisual speech. Proc. of the IEEE 91, 1306–1326. doi:10.1109/JPROC.2003.817150.

Puigcerver, J., Martin-Albo, D., Villegas, M., 2016. Laia: A deep learning toolkit for HTR. URL: https://github.com/jpuigcerver/Laia/.

R Core Team, 2017. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/. Last access: May 2017.

Rabiner, L., Juang, B.H., 1993. Fundamentals of Speech Recognition. Prentice Hall.

Revuelta-Martínez, A., Rodríguez, L., García-Varea, I., 2012. A computer assisted speech transcription system, in: Proc. of the 13th EACL, pp. 41–45.

Romero, V., Sanchez, J.A., Bosch, V., Depuydt, K., de Does, J., 2015. Influence of text line segmentation in handwritten text recognition, in: Document Analysis and Recognition (ICDAR), 2015 13th International Conference on, IEEE. pp. 536–540.

Romero, V., Toselli, A.H., Vidal, E., 2009. Using Mouse Feedback in Computer Assisted Transcription of Handwritten Text Images, in: Proc. of ICDAR '09, pp. 96–100.

Romero, V., Toselli, A.H., Vidal, E., 2012. Multimodal Interactive Handwritten Text Transcription. Series in Machine Perception and Artificial Intelligence (MPAI), World Scientific Publishing.

Sebe, N., Cohen, I., Huang, T.S., 2005. Multimodal emotion recognition. Handbook of Pattern Recognition and Computer Vision 4, 387–419.

Serrano, N., Castro, F., Juan, A., 2010. The RODRIGO Database, in: Proc. of the 7th LREC, pp. 2709–2712. URL: http://aclweb.org/anthology/L10-1330.

Silvestre-Cerdà, J.A., Pérez, A., Jiménez, M., Turro, C., Juan, A., Civera, J., 2013. A System Architecture to Support Cost-Effective Transcription and Translation of Large Video Lecture Repositories, in: Proc. of SMC '13, pp. 3994–3999. doi:10.1109/SMC.2013.682.

Stolcke, A., 2002. SRILM-an extensible language modeling toolkit., in: Proc. of the 3rd Interspeech, pp. 901–904.

Stolcke, A., Bratt, H., Butzberger, J., Franco, H., Gadde, V.R.R., Plauché, M., Richey, C., Shriberg, E., Sönmez, K., Weng, F., Zheng, J., 2000. The SRI March 2000 Hub-5 conversational speech transcription system, in: Proc. of NIST Speech Transcription Workshop.

Stolcke, A., Konig, Y., Weintraub, M., 1997. Explicit word error minimization in n-best list rescoring., in: Proc. of the 5th Eurospeech, pp. 163–166.

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

Villegas, M., Romero, V., Sánchez, J.A., 2015. On the modification of binarization algorithms to retain grayscale information for handwritten text recognition, in: Iberian Conference on Pattern Recognition and Image Analysis, Springer. pp. 208–215.

Wang, T., Wu, D.J., Coates, A., Ng, A.Y., 2012. End-to-end text recognition with convolutional neural networks, in: Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012), pp. 3304–3308.

Welch, B.L., 1947. The Generalization of 'Student's' Problem when Several Different Population Variances are Involved. Biometrika 34, 28–35.

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., et al., 2006. The HTK book. Cambridge university engineering department .