Document downloaded from:

http://hdl.handle.net/10251/65730

This paper must be cited as:

Granell Romero, E.; Martínez-Hinarejos, C. (2015). Multimodal output combination for transcribing historical handwritten documents. En Computer Analysis of Images and Patterns. Springer. 246-260. http://hdl.handle.net/10251/65730.

The final publication is available at

http://link.springer.com/chapter/10.1007/978-3-319-23192-1_21

Copyright Springer

Additional Information

The final publication is available at Springer via http://dx.doi.org/10.1007/978-3-319-23192-1_21

Multimodal Output Combination for Transcribing Historical Handwritten Documents

Emilio Granell and Carlos-D. Martínez-Hinarejos

Pattern Recognition and Human Language Technology Research Center,
Universitat Politècnica de València, Camino Vera s/n, 46022, Valencia, Spain
{egranell, cmartine}@dsic.upv.es

Abstract. Transcription of digitalised historical documents is an interesting task in the document analysis area. This transcription can be achieved by using Handwritten Text Recognition (HTR) on digitalised pages or by using Automatic Speech Recognition (ASR) on the dictation of contents. Moreover, another option is using both systems in a multimodal combination to obtain a draft transcription, given that combining the outputs of different recognition systems will generally improve the recognition accuracy. In this work, we present a new combination method based on Confusion Network. We check its effectiveness for transcribing a Spanish historical book. Results on both unimodal combination with different optical (for HTR) and acoustic (for ASR) models, and multimodal combination, show a relative reduction of Word and Character Error Rate of 14.3% and 16.6%, respectively, over the HTR baseline.

Keywords: Document analysis and transcription, handwritten text recognition, automatic speech recognition, Confusion Networks combination, recognition outputs combination

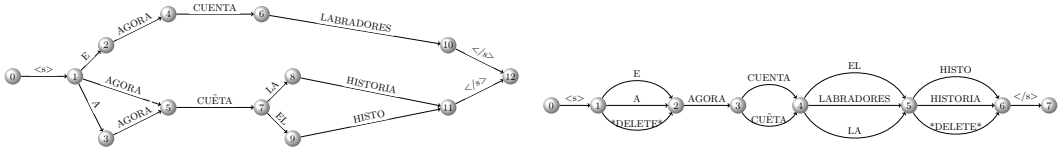
1 Introduction

Document analysis and recognition is a popular field of application of image analysis. This field includes several interesting applications in real tasks, such as the automatic transcription of forms or the transcription of handwritten documents. This last application is of particular interest for recovering the contents of ancient manuscripts available in many historical libraries. Transcribing the handwritten text of these manuscripts would allow to access their contents in a more comfortable manner, which can be interesting for historical, cultural, and even legal purposes.

In this context, transcription of these documents could benefit of the use of image analysis and natural language processing techniques, and more specifically of Handwritten Text Recognition (HTR) techniques [14]. The basic idea is to provide paleographers (expert people in transcribing ancient documents) with an initial draft transcription that they can properly ammend, decreasing the effort in the global transcription task. Moreover, these systems can use feedback from transcribers in order to speed up the transcription process [15]. The high interest in this task within the field of Digital Humanities is reflected in projects such as IMPACT[1] or tranScriptorium[2].

As an alternative to HTR systems, many paleographers asked for dictation systems based on Automatic Speech Recognition (ASR) engines, since many of them claim that it is more comfortable for them to dictate the document contents. Since HTR and ASR systems share most part of the technology related to the recognition process (they are usually based on Hidden Markov Models

[1] http://www.impact-project.eu/
[2] http://transcriptorium.eu



(a) Word Graph lattice. (b) Confusion Network.

Fig. 1. Word Graph lattice and Confusion Network.

-HMM- with N-grams recognition systems using regular or tandem features), the possibility of combining both types of systems arises immediately. This combination would take advantage from two different data sources.

Systems combination is not restricted to different modalities (image and speech) systems, but it can be also applied to the combination of different systems of the same modality (unimodal system combination). In this case, many techniques have been proposed [2,10,7]. In the case of bimodal combination, [18] presents a technique that can be applied for isolated words. However, bimodal combination in continuous decoding is a hard problem because of time asynchrony between the two signals, i.e., the sequence of feature vectors for each modality differs in length and it is not easy to find the time points where the same elements (words in this case) are synchronised. An initial approximation for this case was presented in [1].

In this work, a new technique of recognition output combination is presented. This technique is based on the combination of Confusion Networks [19] derived from the outputs of two recognition systems, of the same or different modality. Furthermore, this technique can be used sequentially to combine the outputs of more than two recognition systems, by using hierarchical combination.

Results show that several errors can be corrected on the resulting unimodal Confusion Network. Nevertheless, the greatest improvements are reached with the multimodal combination. Moreover, results show that the combined Confusion Networks store better quality recognition hypotheses than those of the plain systems. These hypotheses can be used to speed up the transcriber task.

The paper is organised as follows: Section 2 presents the combination technique; Section 3 details the experimental framework (data, conditions, and assessment); Section 4 shows the results of the different experiments; Section 5 offers the final conclusions and future work lines.

2 Confusion Network Combination

As an output format for handwriting and speech recognisers, Confusion Networks (CN) reduce the complexity of Word Graph (WG) lattices without losing important information [19]. Figure 1 provides an example of WG and its corresponding CN. A CN is a weighted directed graph, in which each hypothesis goes through all the nodes. The words and their probabilities are stored in the edges, and the total probability of the words contained in a subnetwork (all edges between two consecutive nodes) sum 1.

The first step of our CN combination technique, aligns by similarity the subnetworks of both CN, marking the selected subnetworks as immovable anchors. In the second step, a new CN is composed from the base of the first CN, by using combination, insertion and deletion of subnetworks.

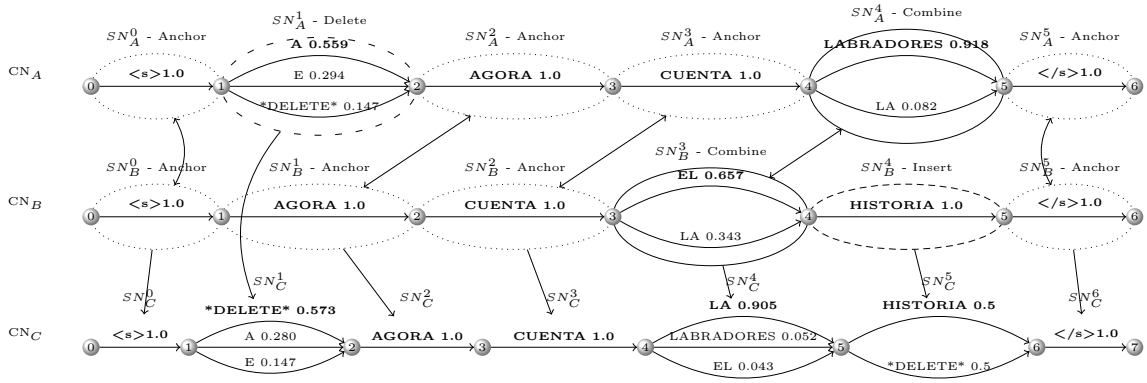


Fig. 2. Bimodal combination example, reference sentence: *<s>AGORA CUENTA LA HISTORIA </s>*.

2.1 Subnetworks based alignment

Due to the fact that outputs of different recognition systems may have different errors, it is necessary to find some reference subnetworks that serve as anchors between the CN to combine. The search of anchor subnetworks is performed in both directions (from left to right and vice versa) simultaneously, taking as anchor subnetworks only those where the search on both directions coincide. This search can be adjusted by several parameters, such as:

- Searching unigrams, bigrams or skip-bigrams.
- Searching only in the most probable word or in all the words contained in the subnetworks.
- Setting a gram matching error threshold ϵ between words.

As words can be decomposed on characters (basic unit for HTR) and phonemes (basic unit for ASR), the quadratic mean of the Character Error Rate (CER) and the Phoneme Error Rate (PER) was used to assess the gram matching error between words of both CN:

$$\mathrm{E}(w_A, w_B) = \sqrt{\frac{\mathrm{CER}(w_A, w_B)^2 + \mathrm{PER}(w_A, w_B)^2}{2}} \tag{1}$$

Where the words of the first and the second CN are represented by w_A and w_B respectively. CER and PER are Levensthein distances between the words of both CN. CER is the distance at character level, and PER at phoneme level, by using the phonetic transcriptions of the recognised words. E represents the gram matching error.

In Figure 2 a complete example of the performance of our technique is shown. In this example, CN_A, CN_B represent the two CN to combine, and CN_C the resulting CN. When searching for anchor subnetworks on bigrams and unigrams with $\epsilon = 0$, it would find the following anchor subnetwork pairs: $SN_A^0 - SN_B^0$, $SN_A^2 - SN_B^1$, $SN_A^3 - SN_B^2$, and $SN_A^5 - SN_B^5$.

2.2 Composing a new Confusion Network

The final goal of the CN combination is to compose a new CN with a higher accuracy than the two original CN. The edit operations used to compose the new CN are: combination, insertion and deletion of subnetworks.

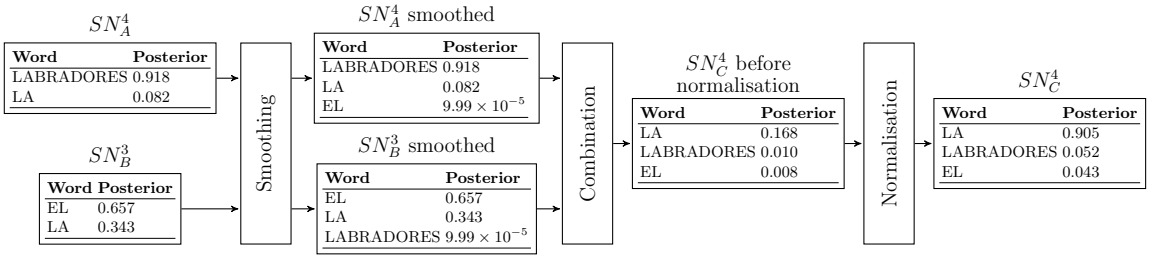


Fig. 3. Example of subnetwork combination with $\alpha = 0.5$ and $\Theta = 10^{-4}$. The Smoothing block represents the use of Equation (4), and the Combination block the use of Equation (3).

Combination Combination of subnetworks allows to maximise the probability of the correct word, if it is present on both subnetworks (SN_A and SN_B). Based on the Bayes theorem and assuming a strong independence on $\Pr(w \mid SN_A, SN_B)$, we get:

$$\Pr(w \mid SN_A, SN_B) \simeq \Pr(w \mid SN_A) \Pr(w \mid SN_B) \tag{2}$$

However, in practice it is usual to employ a weighted version of Equation (2) in which the weight factor α permits to balance the relative reliability between the different probability distributions models (as happens in HTR and ASR systems to balance the influence of the optical/acoustic models and the language model). Thus, in practice we use the following:

$$\Pr(w \mid SN_A, SN_B) \simeq \Pr(w \mid SN_A)^{\alpha} \Pr(w \mid SN_B)^{1-\alpha} \tag{3}$$

Before combining two subnetworks with Equation (3), it is necessary to smooth all the word probabilities. Otherwise, uncommon words would have null probabilities. Equation (4) permits to smooth the word probabilities of all words ($n = common + uncommon$) in both subnetworks. This equation is based on Laplacian smoothing [21]. However, here the word counts are obtained by dividing the word probabilities by a defined granularity Θ.

$$\Pr{}_{smoothed}(w \mid SN) = \frac{\Pr(w \mid SN) + \Theta}{1 + n\Theta} \tag{4}$$

Finally, the word probabilities on the resulting subnetwork are normalised.

In the example (Figure 2), SN_A^4 and SN_B^3 are selected for combination. In this case, the correct word (LA) is not the most probable word in either subnetwork. However, it becomes the most probable word when combining both subnetworks with $\alpha = 0.5$ and $\Theta = 10^{-4}$, as can be seen in SN_C^4 (in Figure 3 this combination process is shown in detail).

Insertion and deletion Insertion and deletion of subnetworks allows to reach a compromise when there is a disagreement between the CN on whether a particular position between two words should or should not be another word. Specifically, insertion occurs when the second CN subnetwork presents a word with a probability greater than a threshold γ and which is not present in the same position of the first CN subnetwork. Deletion occurs similarly, when the second CN considers that a word is not necessary in a specific position of the first CN, and the most probable word of the first CN subnetwork to delete does not reach a threshold δ.

Both operations use the same procedure. First, we halve the probability of all the words of the subnetwork to insert or delete, and then, we increase in a 50% the probability of deleting this subnetwork. As an example, the subnetworks SN_B^4 and SN_A^1 (see Figure 2) are inserted and deleted, respectively.

Composition of the new Confusion Network The first step on the composition of the new CN is to combine the subnetworks labeled as anchor. Thereby between consecutive anchored subnetworks a series of aligned fragments appear in both CN. Each fragment can contain from none to several subnetworks. In the example in Figure 2, two fragments appear, the first one between anchors $SN_A^0 - SN_B^0$ and $SN_A^2 - SN_B^1$, and the second one between anchors $SN_A^3 - SN_B^2$, and $SN_A^5 - SN_B^5$. We use these fragments sizes to decide what to do with each subnetwork. Comparing the sizes of two aligned fragments, we can find the following cases, when both fragment sizes are not null:

1. If both fragment sizes match, all subnetworks are combined one by one.
2. If the fragment size of only one CN is null, we must choose whether to insert or to delete (as explained above) for all the subnetworks contained in the fragment of the other CN. This is the case of the first fragment in the example of Figure 2, which is composed only by SN_A^1. It is deleted, since the probability of the first word does not reach the δ threshold (in this case, $\delta = 0.75$ was chosen).
3. If both fragment sizes are different and none is null, we must find every additional anchor subnetworks in a relaxed search, and decide whether to insert or delete for the rest of subnetworks. This is the case of the second fragment in the example of Figure 2, which is formed by SN_A^4 on a side, and by SN_B^3 and SN_B^4 on the other. When searching for unigrams on the whole subnetworks, it is found that SN_A^4 can be combined with SN_B^3, and given that SN_B^4 exceeds the threshold γ (in this case, $\gamma = 0.25$ was taken), it is inserted.

Finally, a new CN is obtained as a result of this process. In Figure 2, CN_C is the resulting CN; it can be seen that several errors have been corrected, and the correct sentence (*<s>AGORA CUENTA LA HISTORIA </s>*) has the highest probability.

3 Experimental framework

3.1 Corpora

The historical handwritten text corpus used for this work is RODRIGO [16]. It 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 the book is historical chronicles of Spain. Most pages form a single block with well separated lines (usually 25 lines per page), written in calligraphical text. The corpus is available with the lines separated, which are the source of feature extraction (Figure 4 shows an example of a separated line). The corpus has a total of 20,356 lines in PNG format. Some standard partitions were defined for baseline experiments, by using a training set of 5000 lines (about 205 pages). The test set for multimodal experiments is a set of 50 lines (pages 515 and 579). The set of symbols present in this corpus gets modelled by 106 HMM, which take into account lowercase and uppercase letters, numbers, punctuation marks, special symbols, and blank spaces.

For the training of the ASR acoustic models we used a partition of the Spanish phonetic corpus Albayzin [12]. This corpus consists of a set of three sub-corpus recorded by 304 speakers using a

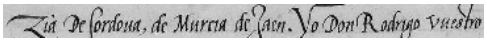


Fig. 4. Example of an extracted line from the RODRIGO corpus.

sampling rate of 16 kHz and a 16-bit quantisation. The training partition used in this work includes 4800 phonetically balanced utterances. Specifically, 200 utterances were read by four speakers and 25 utterances were read by 160 speakers, with a total length of about 4 hours. A set of 25 HMM (23 monophones, short silence, and long silence) was estimated from this corpus. For the multimodal test we recorded 7 different speakers who read the 50 handwritten test lines (those of pages 515 and 579), giving a total set of 350 utterances (about 15 minutes).

3.2 Features

HTR features Handwritten text features are computed in several steps. First, we perform a bright normalisation by using the `pnmnorm` tool of Netpbm [13], with `-bvalue 150 -wvalue 200`. After that, we apply a median filter of size 3×3 pixels to the whole image. Next, we perform slant correction by using the maximum variance method and a threshold of 92%. Then, we perform a size normalisation and scale the final image to a height of 40 pixels. Finally, features are extracted by using the method described in [5]. Final regular feature vectors are composed of 60 dimensions.

ASR features Mel-Frequency Cepstral Coefficients (MFCC) are extracted from the audio files. The Fourier transform is calculated every 10ms 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. We use the first 12 MFCC and the log frame energy with the first and second order derivatives as regular ASR features, resulting in a 39 dimensional vector.

Tandem features In the tandem feature extraction scheme, two Multi-Layer Perceptrons (MLP) with 2000 neurons at the hidden layer and a softmax transfer function at the output layer were trained to estimate symbol-phoneme posterior probabilities. On the one hand, for HTR, a MLP with 60 neurons at the input layer and 106 neurons at the output layer was trained with Torch [4]. On the other hand, for ASR, the MLP had 39 neurons at the input layer and 25 neurons at the output layer and it was trained using QuickNet [8]. The final tandem features are constituted by the log posteriors probabilities of the MLP.

Both MLP's were trained by backpropagation with a mean-squared error criterion. The frame-level labelling required to train the MLP's can be generated from a forced alignment decoding by a previously trained recognition system [6]. For this work, the forced alignment decoding and the model training were repeated several times until the convergence of the frame labels.

3.3 Models

Optical and acoustic models were trained by using HTK [20]. On the one hand, symbols on the optical model are modelled by a continuous density left-to-right HMM with 6 states and 32 gaussians per state. On the other hand, phonemes on the acoustic model are modelled as a left-to-right HMM with 3 states and 64 gaussians per state.

In order to test the influence of the speaker adaptation in the acoustic models, the speaker independent acoustic models were adapted to each speaker and page using HTK's Maximum Likelihood Linear Regression global adaptation [20]. For each independent acoustic model (regular and tandem), two adapted models were obtained per speaker, since we got the adapted models for decoding one page by using the audio samples of the other page, and vice-versa.

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

The language model was estimated directly from the transcriptions of the 5000 lines included in the HTR training set by using the SRILM *ngram-count* tool [17]. The model was a 2-gram with the Kneser-Ney back-off smoothing [9]. The test baseline language model presents 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 book makes it difficult to find representative texts to enhance the language model.

3.4 Evaluation metrics

The Levensthein distance was used to assess our system at the word level with the Word Error Rate (WER) and at the character level with the Character Error Rate (CER). These measures calculate the error between the best hypothesis from decoding processes and the reference. CER is especially interesting within this framework, since transcription errors are usually corrected at character level. For both measures, confidence intervals at 95% were calculated by using the bootstrapping method with 10,000 repetitions [3]. Similarly, oracle WER and oracle CER are the best WER and best CER, respectively, that can be obtained from the hypotheses present in a Confusion Network. In this work, the search of oracle values was limited to the 2000-best.

The statistical dispersion of the position of the n-best values that allow to obtain the oracle permits to estimate the difficulty of obtaining this oracle. The interquartile range (IQR), the median, and the median absolute deviation (MAD) were used to measure this statistical dispersion.

3.5 Experimental setup

The recognition systems were implemented by using the iATROS [11] recogniser, and the SRILM *lattice-tool* [17] utility was used to obtain CN from the n-best WG recognition outputs.

In the experiments, the anchor subnetworks search was performed several times, looking for skip-bigrams (allowing only one wrong word in the gap) and unigrams, throughout the whole hypothesis in the subnetworks. Specifically, we started with a perfect matching search of skip-bigrams, followed by a perfect matching search of unigrams. Next, we made a relaxed matching search of skip-bigrams, and a relaxed matching search of unigrams, setting the matching error threshold to $\epsilon = 2^{-\frac{1}{2}}$. A relaxed search with this threshold would allow to align words like *3* and *TRES* that coincide in their phonetic transcription (['tres]), since it is the same word written differently.

4 Experimental results

In order to check the performance of our technique we used six different recognition systems; two HTR systems (regular and tandem), and four ASR systems: regular and tandem, with (A) and without speaker adaptation. As a first step, the reference values for each recognition system were obtained. As a second step, the unimodal combination was performed. Finally, the multimodal

Table 1. *HTR baseline results, along with the average decoding times for each sample (in seconds).*

Model	1-best WER	Oracle WER	1-best CER	Oracle CER	Time per sample
Regular	39.3% ± 4.1	28.0% ± 3.1	20.6% ± 2.5	14.0% ± 2.1	36.7 s
Tandem	**32.9% ± 5.3**	**24.9% ± 5.6**	**15.7% ± 4.9**	**11.6% ± 4.9**	**15.4 s**

Table 2. *ASR baseline results, along with the average decoding times for each sample (in seconds).*

Model	1-best WER	Oracle WER	1-best CER	Oracle CER	Time per sample
Regular	62.9% ± 2.2	44.9% ± 2.1	35.4% ± 1.4	25.1% ± 1.4	55.2 s
Regular (A)	**51.0% ± 2.2**	**33.7% ± 2.0**	**26.4% ± 1.3**	**17.2% ± 1.2**	43.2 s
Tandem	58.6% ± 2.2	41.1% ± 2.0	31.3% ± 1.3	21.9% ± 1.3	30.7 s
Tandem (A)	55.5% ± 2.2	38.4% ± 2.1	28.9% ± 1.3	19.7% ± 1.1	**29.4 s**

combination was conducted. With regard to the order of the combinations, the best results were obtained with the order represented in Figure 5. However, the differences were not significant with respect to the rest of possible combination orders.

The values of the main variables were tuned, in order to optimise the experimental results. We used a weight factor of $\alpha = 0.5$, a granularity for smoothing of $\Theta = 10^{-4}$, and $\gamma = 0.25$ and $\delta = 0.75$ as thresholds for insertion and deletion, respectively.

4.1 Baseline experiments

The reference values obtained for each one of the six recognition systems are shown in Table 1 and Table 2. As can be observed, the HTR reference values are better than the ASR reference values.

As to the baseline HTR results (Table 1), the tandem system produces lower error rates. Therefore, we take these values as a baseline reference, namely 32.9% for WER and 15.7% for CER. Regarding the oracle baseline results, the lower bounds were obtained in the HTR modality, specifically 24.9% for WER and 11.6% for CER. Both are also from the tandem system.

The baseline ASR results (Table 2) are quite poor because of the difficulty of the corpus. In RODRIGO we faced with text images containing hyphenated words (e.g., *REYNA*, where a part of the word *RE* is at the end of a line and the second part *YNA* is at the beginning of the following line), abbreviations (e.g., *NRÕ*) that are pronounced as the whole word (*NUESTRO* ['nwes tro]), and words written in multiple forms (e.g., *XPIÃNOS* and *CHRISTIANOS*, or numbers as *5* and *V*) but that are pronounced in the same way ([kris 'tja nos], ['θiŋ ko]). In spite of these facts, speaker adaptation and tandem features provide improvements for ASR when compared to the regular baseline. Specifically, the best results were obtained with the regular model with speaker adaptation (A). Regarding the WER, a value of 51.0% was obtained, while the value of CER reached 26.4%. In terms of the oracle, the ASR oracle values are worse than the HTR baseline values.

Concerning the average decoding times per sample, we take as a reference 55.2 s, which is the time of the slower decoding system (ASR regular), because it represents the average time required to decode a sample by all the systems in parallel.

4.2 Unimodal combination experiments

In these experiments, the output of the different systems of the same modality were combined. The input signals to each recognition system are the same. Therefore, the combination of these outputs does not present the difficulty of asynchrony.

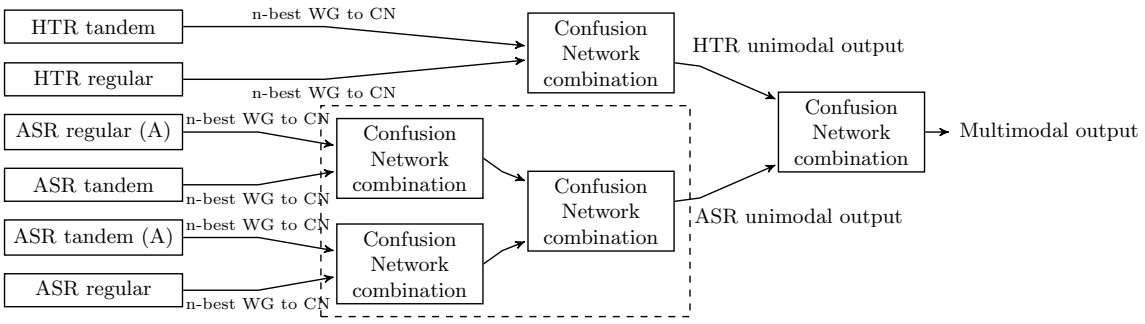


Fig. 5. Unimodal - Multimodal combination diagram.

On the one hand, in the case of the HTR unimodal combination (Figure 5), once the image signals were processed through each HTR system, the n-best WG outputs were transformed into CN. Then, these CN were processed by our CN combination technique, which returned a new CN by combining the information from both HTR systems.

On the other hand, for the ASR unimodal combination, as our technique allows to combine only two CN and we used four ASR systems, it was necessary to use the CN combination technique three times as described in Figure 5. First, the voice signals were processed through each ASR system, and its n-best WG outputs were transformed in CN. Secondly, these CN were processed by our CN combination technique by pairs, in order to obtain two combined CN. Finally, the two combined CN were processed by our CN combination technique, thereby we got a new CN that combines the information from the four ASR systems.

As shown in Table 3, the HTR unimodal combination reduced the error with respect to the baseline HTR system at both the WER and CER level, 1.8% and 2.4% respectively. Meanwhile, the combination introduced new information in the new CN that reduced the oracle levels, being of special interest the case of the oracle CER since it presents 32.8% of relative reduction over the HTR oracle CER baseline reference. In the case of the ASR unimodal combination, only a small improvement is produced at the WER level when compared to the ASR baseline.

The time required for combining two CN is related to its density. The average time per sample used in the unimodal combinations was 457.1 ms for HTR and 596.9 ms for ASR. Nevertheless, due to the two combination levels for ASR, the actual average time used for ASR was 1.2 s.

4.3 Multimodal combination experiment

In the multimodal combination experiment, the unimodal CN were combined with the aim of obtaining a multimodal CN (Figure 5). The multimodal CN combination produced improvements compared with the results obtained by the unimodal combinations; despite the high error values obtained by the ASR unimodal combination. As can be seen in Table 3, a relative WER improvement of 9.3% is achieved when compared to the HTR unimodal combination, and this relative improvement increases to 14.3% when compared to the baseline reference as well. Furthermore, in terms of the CER a relative improvement of 1.5% is produced when compared to the HTR unimodal combination; besides, the relative improvement over the baseline reached 16.6%.

Table 3. *Combination results, along with the average combination times for each pair of samples (in miliseconds).*

Combination	1-best WER	Oracle WER	1-best CER	Oracle CER	Time per sample
HTR unimodal	$31.1\% \pm 3.7$	$20.6\% \pm 3.2$	$13.3\% \pm 1.9$	$7.8\% \pm 1.4$	457.1 ms
ASR unimodal	$50.4\% \pm 2.0$	$34.9\% \pm 1.9$	$28.6\% \pm 1.3$	$18.5\% \pm 1.2$	596.9 ms
Multimodal	$\mathbf{28.2\% \pm 1.3}$	$\mathbf{16.6\% \pm 1.3}$	$\mathbf{13.1\% \pm 0.7}$	$\mathbf{6.2\% \pm 0.5}$	596.2 ms

The oracle WER level presents a statistically significant relative improvement of 33.3% when compared to the oracle WER baseline, while the statistically significant relative improvement presented by the oracle CER level when compared with the oracle CER baseline attained 46.6%.

The average time per sample was 596.2 ms, considering the necessary the time to obtain the unimodal combinations this value increases to 1.8 s, which means that the required time for the multimodal combination represents 3.3% of relative increase of extra time to the decoding processes.

4.4 Difficulty of reaching the oracle values

The difficulty of reaching the oracle values from the information contained in each confusion network was estimated by means of a statistical study of the dispersion of the n-best positions that allowed to achieve those oracle values.

In Table 4 the statistical dispersion obtained for each CN is outlined, for oracle WER and oracle CER. Regarding the CN obtained from the recognition systems, the CN obtained from the tandem model of the HTR offered the narrowest IQR, side by side with the lowest median and MAD values. On the opposite side, all the CN of the ASR presented a wide IQR with high median and MAD values. As to the CN acquired from the combinations, the depth of the search is reduced because of the increase in the amount of information and the word probability correction resulting from the combination.

To show the importance of these statistical dispersions it is necessary to represent them with their related error values, as in Figure 6. In Figures 6(a) and 6(b), the statistical dispersions are plotted as box plots, where the positions are normalised, the range of values is set to the difference between the 1-best error value and the oracle error value, and the minimum value is set to the oracle error value. In the box plots, the IQR, the median, the minimum and the maximum values of the statistical dispersions are represented. The smaller the box plot, the easier it will be to reach the oracle error value. Therefore, the better systems present the lower values.

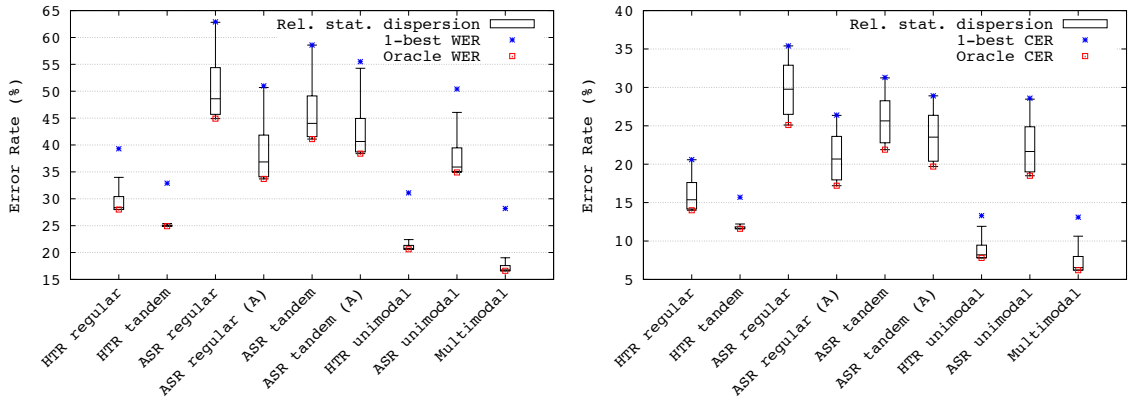
Regarding the oracle WER (Figure 6(a)), in the CN of the reference system (HTR tandem) it is very easy to reach the oracle WER value (24.9%) from the WER value (32.9%), whereas in the CN from the different ASR systems it is not easy to reach the oracle WER value. With the combination, the difficulty of reaching the oracle values is reduced, especially in the CN obtained from the multimodal combination where it is quite easy to reach the oracle WER value (16.6%) from the WER value (28.2%). In the oracle CER (Figure 6(b)), a similar behaviour is observed.

5 Conclusions

In this study, we have confirmed the benefits of combining multiple recognition outputs for the transcription of handwritten historical text. The technique presented in the paper takes advantage

Table 4. *Statistical dispersion of the n-best positions that allow to obtain the oracle values.*

Confusion network	Oracle WER			Oracle CER		
	IQR	Median	MAD	IQR	Median	MAD
HTR regular	423	62	61	1041	415	409
HTR tandem	50	2	2	119	17	16
ASR regular	967	412	398	1242	906	617
ASR regular (A)	897	365	359	1230	756	601
ASR tandem	866	334	329	1164	796	591
ASR tandem (A)	728	263	260	1300	831	658
HTR unimodal combination	137	26	25	594	144	142
ASR unimodal combination	567	129	127	1161	625	573
Multimodal combination	166	29	28	510	100	99



(a) Statistical dispersion relative to 1-best WER and oracle WER.

(b) Statistical dispersion relative to 1-best CER and oracle CER.

Fig. 6. Relative statistical dispersion in the set of the positions of the n-best that obtain the oracle values.

of the fact that different systems make different errors; thus, editing operations can correct errors. Insertion and deletion create new bigrams than enrich the resulting CN, and the combination can maximise the probability of the correct word, when both subnetworks contain the correct word, even when this word has a low probability in both subnetworks. Conversely, if only one subnetwork contains the correct word and both subnetworks contain the same erroneous word, this error will be maximised at the expense of the correct word. Despite of this fact, the experiments performed confirm the strengths of this CN combination technique.

The obtained results show that there is still room for improvement. For example, one of the improvements that could be made is to extend the technique in order to combine the outputs of more than two recognition systems simultaneously. Moreover, we propose for future studies the use of more robust methods of optical and acoustic modelling, such as the use of Recurrent Neural Networks with Long Short-Term Memory features on the tandem approach. From the ASR point of view, the possibility of using not lines but whole sentences of the handwritten text corpus could make multimodality more natural. Eventually, integrating the technique presented in this paper in an interactive transcription system could reduce the time and the workload for transcribing

historical books, due to the increased recognition accuracy and the facility of achieving the oracle values on the resulting CN, without a significant increase of processing time.

Acknowledgements Work partially supported by European Union - 7th FP, under grant 600707 (tranScriptorium), and by the Spanish MEC under projects STraDA (TIN2012-37475-C02-01), Active2Trans (TIN2012-31723), and SmartWays (RTC-2014-1466-4).

References

1. Alabau, V., Martínez-Hinarejos, C.D., Romero, V., Lagarda, A.L.: An iterative multimodal framework for the transcription of handwritten historical documents. Pattern Recognition Letters 35, 195–203 (2014)
2. Bertolami, R., Halter, B., Bunke, H.: Combination of multiple handwritten text line recognition systems with a recursive approach. In: Proc. Int. Conf. Frontiers Handwriting Recognition. pp. 61–65 (2006)
3. Bisani, M., Ney, H.: Bootstrap estimates for confidence intervals in ASR performance evaluation. In: Proc. of Int. Conf. on Acoustics, Speech and Signal Processing. vol. 1, pp. 409–412 (2004)
4. Collobert, R., Bengio, S., Mariéthoz, J.: Torch: a modular machine learning software library. Tech. rep., IDIAP-RR 02-46, IDIAP (2002)
5. Dreuw, P., Jonas, S., Ney, H.: White-space models for offline Arabic handwriting recognition. In: Proc. of Int. Conf. on Pattern Recognition. pp. 1–4 (2008)
6. Hermansky, H., Ellis, D.P., Sharma, S.: Tandem connectionist feature extraction for conventional HMM systems. In: Proc. of Int. Conf. Acoustics, Speech and Signal Processing. vol. 3, pp. 1635–1638 (2000)
7. Ishimaru, S., Nishizaki, H., Sekiguchi, Y.: Effect of Confusion Network Combination on Speech Recognition System for Editing. In: Proc. of APSIPA Annual Summit and Conf. vol. 4, pp. 1–4 (2011)
8. Johnson, D.: ICSI Quicknet soft package (2004), `http://www1.icsi.berkeley.edu/Speech/qn.html`
9. Kneser, R., Ney, H.: Improved backing-off for m-gram language modeling. In: Proc. of Int. Conf. Acoustics, Speech and Signal Processing. vol. 1, pp. 181–184 (1995)
10. Krishnamurthy, H.K.: Study of algorithms to combine multiple automatic speech recognition (ASR) system outputs. Master's thesis, Department of Electrical and Computer Engineering (2009), `http://hdl.handle.net/2047/d10019273`
11. Luján-Mares, M., Tamarit, V., Alabau, V., Martínez-Hinarejos, C.D., Pastor i Gadea, M., Sanchis, A., Toselli, A.H.: iATROS: A speech and handwritting recognition system. In: V Jornadas en Tecnologías del Habla (VJTH2008). pp. 75–78 (2008)
12. Moreno, A., Poch, D., Bonafonte, A., Lleida, E., Llisterri, J., Mariño, J.B., Nadeu, C.: Albayzin speech database: Design of the phonetic corpus. In: Proc. of EuroSpeech'93. pp. 175–178 (1993)
13. Netpbm home page, `http://netpbm.sourceforge.net/`
14. Plamondon, R., Srihari, S.N.: On-Line and Off-Line Handwriting Recognition: A Comprehensive Survey. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(1), 63–84 (January 2000)
15. Romero, V., Leiva, L.A., Toselli, A.H., Vidal, E.: Interactive Multimodal Transcription of Text Images Using a Web-based Demo System. In: Proc. of Conf. on Intelligent User Interfaces. pp. 477–478 (2009)
16. Serrano, N., Castro, F., Juan, A.: The RODRIGO Database. In: Proc. of Language Resources and Evaluation Conference. pp. 2709–2712 (2010)
17. Stolcke, A.: SRILM - An extensible language modeling toolkit. In: Proc. Interspeech. pp. 901–904 (2002)
18. Woodruff, P., Dupont, S.: Bimodal Combination of Speech and Handwriting for Improved Word Recognition. In: Proc. of EUSIPCO 2005. pp. 1918 – 1921 (2005)
19. Xue, J., Zhao, Y.: Improved confusion network algorithm and shortest path search from word lattice. In: Proc. of Int. Conf. in Acoustics, Speech and Signal Processing. vol. 1, pp. 853–856 (2005)
20. Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., et al.: The HTK book (for HTK version 3.4). Cambridge university eng. dept. (2006)
21. Zhai, C., Lafferty, J.: A study of smoothing methods for language models applied to information retrieval. Transactions on Information Systems 22(2), 179–214 (2004)