COMPUTER-ASSISTED HANDWRITTEN TEXT TRANSCRIPTION USING SPEECH RECOGNITION

Antonio-L. Lagarda, Vicent Alabau, Carlos-D. Martínez-Hinarejos,
Alejandro-H. Toselli, Verónica Romero, José-R. Navarro and Enrique Vidal

Instituto Tecnológico de Informática
Universidad Politécnica de Valencia
Camino de Vera, s/n, 46022 Valencia, Spain
{alagarda, valabau, cmartine, ahector, vromero, jonacer, evidal}@iti.upv.es

ABSTRACT

Handwritten Text Recognition (HTR) has gained a lot of attention during the last few years. On the one hand, there exists a large number of applications that can benefit from it. On the other hand, HTR is similar to Automatic Speech Recognition (ASR). Thus, modelling techniques can be easily adapted and most of the already available ASR tools can be used to train and evaluate HTR systems quite straightforwardly. However, HTR performance is still far from perfect and post-editing is needed. This post-editing can be made efficiently by means of Computer Assisted Transcription of Handwritten Text Images (CATTI) tools, which iteratively interact with the user to achieve the desired transcription. Typically, this interaction has been made via mouse and keyboard. In this paper, we present the development of a new CATTI system which uses speech as an additional mean of interaction. The system is build upon a generic recogniser both for speech and handwriting recognition.

1. INTRODUCTION

Handwriting recognition has become an interesting task in the last years due to its multiple applications. These applications cover from the use of mobile devices where handwritten input is used [1] to the transcription of handwritten documents, specially of ancient documents with high historic value [2].

The recent success of Handwritten Text Recognition (HTR) is based on the use of models that were previously used in other tasks such as speech recognition. Nowadays, handwritting recognition systems make use of Hidden Markov Models (HMM) to define the morphological features of the handwritten text [3]. They also use language models (usually N-grams) to define the relations between the words to be recognised.

However, recognition accuracy is far to be perfect, and post-editing is usually needed. Apart from the classical post-editing techniques, based on keyboard and mouse input, a novel approach is the use of speech to validate the correctly recognised text. This approach has been previously used in other computer-assisted task, such as Computer-Assisted Translation (CAT) [4], with promising results. Moreover, the validation of the correct recognition by the user entails the possibility of using this information as feedback to improve the recognition models.

In this paper, we show the development of a Computer-Assisted Transcription of Handwritten Text Images (CATTI) [2] system which uses validation based on speech input, i.e., a multimodal system. The system is build using a generic recogniser, the iATROS toolkit [5], which allows an easy construction of multimodal applications. The iATROS toolkit can perform handwriting and speech recognition. Both types of recognitions use models of the same nature and the same search process.

The paper is organised as follows: Section 2 presents the theoretical foundations of the CATTI process and the integration of the speech input in the CATTI framework. Section 3 describes the use of our CATTI system. Section 4 specifies the architecture and relations between the different modules that have been used in the construction of our system. Section 5 presents some conclusions and future improvements for the system.

2. FRAMEWORK

This section introduces the theoretical framework of the CATTI process and how speech input can be incorporated.

2.1. Handwritten text recognition

The traditional handwritten text recognition problem can be formulated as the problem of finding the most likely word sequence, $\hat{w}$, for a given handwritten sentence image represented by a feature vector sequence x, i. e., $\hat{w} = \arg\max_w \Pr(w|x)$. Using the Bayes' rule we can decompose the probability $\Pr(w|x)$ into two probabilities,

Work supported by VIDI-UPV under project 20070315, by the EC (FEDER) and the Spanish MEC under grant TIN2006-15694-CO2-01, by the Spanish research programme Consolider Ingenio 2010: MIPRCV (CSD2007-00018), and by the i3media Cenit project (CDTI 2007-1012).

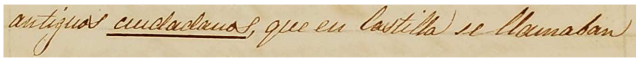
	x							
INTER-0	p							
INTER-1	$\hat{s}$	*antiguos*	*cuidadores*	*que*	*en*	*el castillo*	*sus*	*llamadas*
	v							
	$\hat{a}$	antiguos						
	p	antiguos						
INTER-2	$\hat{s}$		*ciudadanos*	*que*	*en*	*el castillo*	*sus*	*llamadas*
	v							
	$\hat{a}$				en			
	p	antiguos	ciudadanos	que	en			
FINAL	$\hat{s}$					*Castilla*	*se*	*llamaban*
	v							
	$\hat{a}$							#
	$p \equiv w$	**antiguos**	ciudadanos	que	**en**	Castilla	se	llamaban

Figure 1. Example of speech interaction with a computer-assisted transcription of handwritten text images. Given a handwritten text image x, the system finds its most likely transcription $\hat{s}$. Then, the user validates its longest error-free prefix p by uttering (v) the last correct word ($\hat{a}$). Next, the new prefix p is taken by the system to suggest a new improved hypothesis $\hat{s}$. This loop is repeated until a transcription is deemed satisfactory by the user, indicated by "#" in the figure. System suggestions are printed in italics, text decoded from user speech in boldface. In the final translation w, words uttered by the user are those that appear in underlined boldface.

$\Pr(x|w)$ and $\Pr(w)$, representing morphological-lexical knowledge and syntactic knowledge, respectively:

$$\hat{w} = \arg\max_{w} \Pr(w|x) = \arg\max_{w} \Pr(x|w) \cdot \Pr(w) \quad (1)$$

$\Pr(x|w)$ is typically approximated by concatenated character models (usually hidden Markov models [6, 7]) and $\Pr(w)$ is approximated by a word language model (usually n-grams [6]).

In the CATTI framework [2, 8], in addition to the given feature sequence, x, a prefix p of the transcription (validated and/or corrected by the user) is available and the HTR should try to complete this prefix by searching for a most likely suffix $\hat{s}$ as:

$$\begin{aligned}\hat{s} &= \arg\max_{s} \Pr(s|x,p) \\ &= \arg\max_{s} \Pr(x|p,s) \cdot \Pr(s|p) \; . \quad (2)\end{aligned}$$

Equation (2) is very similar to (1), being w the concatenation of p and s. The main difference is that now p is given. Therefore, the search must be performed over all possible suffixes s following p and the language model probability $\Pr(s|p)$ must account for the words that can be written after the prefix p.

2.2. Interaction with speech

In the scenario described in the previous section, the user interacts with the system to achieve a better transcription of the original text. Typically, the user validates and/or corrects the prefix with the keyboard and the mouse. However, this interaction would be more natural and efficient if the user could interact with the speech as well. For that reason, we present an alternative way for the user to interact with the system by means of a speech recogniser in a similar fashion of what was proposed in [4].

The problem of speech recognition can be formulated in a similar way to the problem of handwritten text recognition in Equation (1). Let be v the vector of features representing the speech signal, a a sequence of words indicating the action to be performed and $p(v|a,\hat{w})$ the phonological-lexical knowledge. $p(a,\hat{w})$ is a language model of the possible actions that are understood by the system (usually a finite state machine). This model is constrained to the current iteration hypothesis and to the limited set of actions that are allowed at this iteration. As a result, the problem of obtaining the most likely word sequence of the action, $\hat{a}$, given the speech utterance v from the user and the current most likely transcription $\hat{w}$ can be written as:

$$\begin{aligned}\hat{a} &= \arg\max_{a} \Pr(a|v,\hat{w}) \\ &= \arg\max_{a} \Pr(v|a,\hat{w}) \cdot \Pr(a,\hat{w}) \; . \quad (3)\end{aligned}$$

3. SPEECH INTERACTION WITH A COMPUTER-ASSISTED TRANSCRIPTION SYSTEM

This section details the use of a CATTI system with speech interaction, illustrated in Figure 1. The described CATTI system is similar to that explained in [9]. Given a handwritten text image, the iterative process starts when the HTR module proposes a full transcription $\hat{s}$ of the feature vectors sequence x, extracted from the image. Then, the human transcriptor (namely the user of our system) checks this hypothesis until he or she finds a mistake. In this way, he or she validates a prefix p of the transcription which is error-free. In our case, this validation can be carried out by means of the keyboard or speech: in the former, by moving the cursor until the last correct word; in the latter, just by uttering that word ($\hat{a}$). In this way, the user is not only validating an error-free prefix from

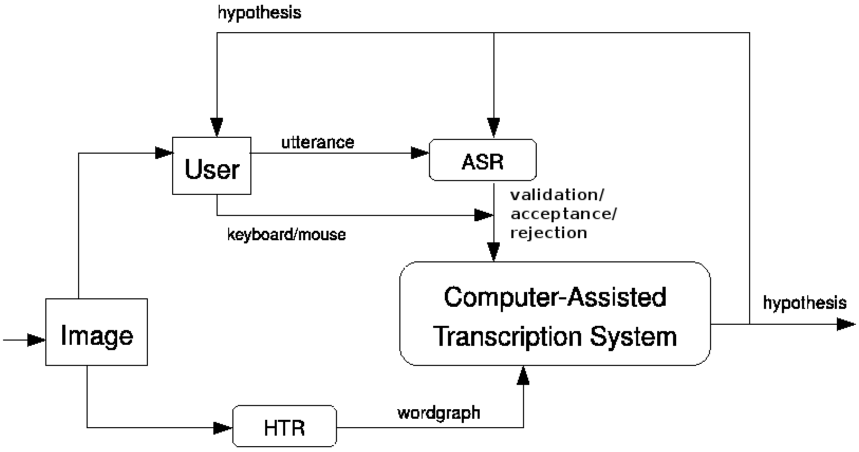


Figure 2. System architecture. From a given handwritten text image, a HTR system (which involves preprocessing, feature extraction, and recognition) produces a word graph of possible transcriptions. This word graph is taken by the Computer-Assisted Transcription System, which starts an iterative process where the user interacts with the system to achieve an error-free transcription. That interaction can be performed by means of speech, the keyboard or the mouse. If speech is employed, the ASR module is in charge of its acquisition, preprocessing and recognition. For further details on this process, see section 3.

the transcription, but also indicating to the system that the following word in the hypothesis is wrong. With that information, the system can suggest an new appropriate suffix $\hat{s}$ to continue that prefix. This cycle is repeated until a correct transcription w of x is accepted by the user by typing an acceptance key sequence or by uttering a reserved acceptance word (in our case, “accept”). There is another special situation in which the first word of the suffix is wrong. In this case, the user just needs to utter a reserved rejection word, and the system will propose a new suffix beginning from a different word.

4. SYSTEM ARCHITECTURE

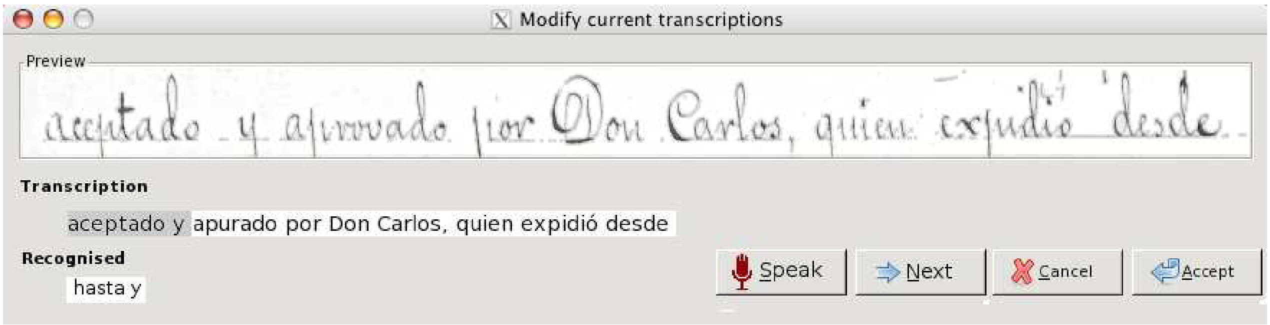
Figure 2 shows a diagram of the system architecture, while Figure 3 is a screen capture of the graphical interface of the application which implements this architecture. Our system is composed of three main components: a HTR system, a CATTI system, and an Automatic Speech Recognition (ASR) system. Given a handwritten text image, the HTR system obtains a word graph with the most likely transcriptions. This word graph is employed by the CATTI system to perform the decoding process, looking for the most probable transcription taking into account the current validated prefix and applying error-correcting techniques when the current prefix is not in the word graph [10]. The CATTI module interacts with the user until the final transcription is achieved (see Section 3 for more details). That interaction can be carried out by means of speech, which is recognized by the ASR system.

Both the HTR and the ASR systems have been implemented based on the iATROS toolkit [5]. Their structure is similar: given an input (respectively, a handwritten text image and an audio signal) the iATROS toolkit is in charge of its preprocessing, the extraction of the feature vectors, and a search based on Viterbi. In both cases, models and search are of the same nature. HTR preprocesses the image to filter out noise, recover handwritten strokes from degraded images and reduce variability of text styles [11]; then, it extracts the image features, where a feature vector sequence is obtained as the representation of the handwritten text image; and finally, the iATROS recognition module obtains the most likely word sequences for the feature vectors in form of a word graph. ASR acquires an utterance from the user, which is preprocessed, and a final recognition is obtained. As seen in section 3, this speech interaction is used for validation purposes, so the language model must be built from the current suffix and the validation/acceptance/rejection reserved words. The small size of this language model ensures a nearly perfect speech recognition.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we have presented the development of a Computer-Assisted Transcription of Handwritten Text Images system in which speech was employed to perform the user-machine interaction.

In this way, the decoding of a human transcriptor utterance is used to validate a prefix of the final transcription. This prefix is taken into account by the CATTI system to suggest a new suffix. The human transcriptor can


Modify current transcriptions
Preview
aceptado y aprovado por Don Carlos, quien expidio desde
Speak
Next
Cancel
Accept

Transcription

aceptado y apurado por Don Carlos, quien expidió desde

Recognised

hasta y

Figure 3. Interface. The original handwritten image appears in the *Preview* box. After an initial transcription is proposed by the system, the user can select a correct prefix and ask for the suffix that best completes it. In the example, once read the suggested transcription, the user decides that its longest error-free prefix is *aceptado y*. Then, he or she pushes the *Speak* button to begin to speak, and utters *hasta y* to select the last correct word. The system will take the new prefix *aceptado y* and try to complete it by suggesting a new suffix.

then accept it or partially validate it by means of speech or by typing in an iterative way until a satisfactory, correct transcription is finally produced.

The iATROS toolkit has been succesfully employed to implement the Handwritten Text Recognition and the Automatic Speech Recognition modules.

In future versions of our system, speaker adaptation techniques could be easily introduced [12], which could improve the speech recognition quality.

In addition, our system could take profit of the user amendments to improve the models quality. In this way, each new iteration, the necessary effort by the user to achieve the correct transcription would be lower, which would enhance both the ergonomics of the system and the user's productivity.

6. REFERENCES

[1] J. Hannuksela, P. Sangi, y J. Heikkila, "Motion-based handwriting recognition for mobile interaction," in *ICPR '06: Proceedings of the 18th International Conference on Pattern Recognition*, Washington, DC, USA, 2006, pp. 397–400, IEEE Computer Society.

[2] V. Romero, A. H. Toselli, L. Rodríguez, y E. Vidal, "Computer Assisted Transcription for Ancient Text Images," in *International Conference on Image Analysis and Recognition (ICIAR 2007)*, vol. 4633 of *LNCS*, pp. 1182–1193. Springer-Verlag, Montreal (Canada), August 2007.

[3] M. Gilloux, *Hidden markov models in handwriting recognition*, vol. 126 of *NATO ASI Series*, Springer Verlag, France, 1994.

[4] E. Vidal, F. Casacuberta, L. Rodríguez, J. Civera, y C. Martínez, "Computer-assisted translation using speech recognition," *IEEE Transaction on Audio, Speech and Language Processing*, vol. 14, no. 3, pp. 941–951, 2006.

[5] M. Luján, V. Tamarit, V. Alabau, C.-D. Martínez-Hinarejos, M. Pastor, A. Sanchis, y A. Toselli, "iATROS: A speech and handwriting recognition system.," in *V Jornadas en Tecnología del Habla*, p. Accepted. Bilbao, 2008.

[6] F. Jelinek, *Statistical Methods for Speech Recognition*, MIT Press, 1998.

[7] L. Rabiner, "A Tutorial of Hidden Markov Models and Selected Application in Speech Recognition," *Proc. IEEE*, vol. 77, pp. 257–286, 1989.

[8] A. H. Toselli, V. Romero, L. Rodríguez, y E. Vidal, "Computer Assisted Transcription of Handwritten Text," in *9th International Conference on Document Analysis and Recognition (ICDAR 2007)*, pp. 944–948. IEEE Computer Society, Curitiba, Paraná (Brazil), September 2007.

[9] V. Romero, A. H. Toselli, J. Civera, y E. Vidal, "Improvements in the computer assisted transcription system of handwritten text images," in *PRIS*, 2008, pp. 103–112.

[10] S. Barrachina, O. Bender, F. Casacuberta, J. Civera, E. Cubel, S. Khadivi, A. Lagarda, H. Ney, J. Tomás, y E. Vidal, "Statistical approaches to computer-assisted translation," *Computational Linguistics*, p. In press, 2008.

[11] A. H. Toselli, A. Juan, J. González, I. Salvador, E. Vidal, F. Casacuberta, D. Keysers, y H. Ney, "Integrated handwriting recognition and interpretation using finite-state models," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 18, no. 4, pp. 519–539, 2004.

[12] P. C. Woodland, "Speaker adaptation for continuous density hmms: A review.," *ITRW on Adaptation Methods for Speech Recognition*, pp. 11–19., August 29-30, 2001.