Pattern Recognition Letters 35 (2014) 195–203

Contents lists available at SciVerse ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

An iterative multimodal framework for the transcription of handwritten historical documents

Vicent Alabau, Carlos-D. Martínez-Hinarejos *, Verónica Romero, Antonio-L. Lagarda

Institut Tecnològic d'Informàtica, Universitat Politècnica de València, Camí de Vera s/n, 46022 València, Spain

ARTICLE INFO

Article history:
Available online 29 November 2012

Keywords:
Ancient text transcription
Handwritten text recognition
Speech dictation
Multimodal systems
Iterative systems
Language modelling

ABSTRACT

The transcription of historical documents is one of the most interesting tasks in which Handwritten Text Recognition can be applied, due to its interest in humanities research. One alternative for transcribing the ancient manuscripts is the use of speech dictation by using Automatic Speech Recognition techniques. In the two alternatives similar models (Hidden Markov Models and *n*-grams) and decoding processes (Viterbi decoding) are employed, which allows a possible combination of the two modalities with little difficulties. In this work, we explore the possibility of using recognition results of one modality to restrict the decoding process of the other modality, and apply this process iteratively. Results of these multimodal iterative alternatives are significantly better than the baseline uni-modal systems and better than the non-iterative alternatives.

1. Introduction

In the last years, many on-line archives and digital libraries are publishing large quantities of digitised legacy documents. These documents must be transcribed into an appropriate textual electronic format in order to allow text-based search of their contents and provide historians and other researchers new ways of indexing, consulting and querying their contents. However, the vast majority of these documents (hundreds of terabytes of digital image data) remain waiting to be transcribed into a textual electronic format. Therefore, manual transcription of these documents is an important task for making available the contents of digital libraries.

These transcriptions are usually carried out by experts in paleography, who are specialised in reading ancient scripts. These scripts are characterised by different handwritten/printed styles from diverse places and time periods. The time that takes for an expert to make a transcription of one of these documents depends on their skills and experience. Most paleographs agree that each page needs several hours to be transcribed.

In this context, Handwritten Text Recognition (HTR) (Marti and Bunke, 2001;Toselli et al., 2004;Plötz and Fink, 2009) has become an important research topic. HTR tries to obtain the word sequence contained in the image of a handwritten text line. This process needs a previous detection of lines of text in an image, as well as some preprocessing steps to make the handwritten text more regular. The final result is a sequence of words (transcription) of the text line, that may contain errors. When the rate of errors of the transcription is low enough, HTR can be a very useful tool to speed up the transcription of handwritten text documents.

However, when consulting paleographs on the most comfortable method to transcribe a handwritten text document, many of them claim that a dictation of the words is the best option. Consequently, Automatic Speech Recognition (ASR) systems are an important alternative to HTR systems. In addition, the current state-of-the-art ASR and HTR systems share many features: Hidden Markov Models (HMM) (Jelinek, 1998;Rabiner, 1989) are used to model the basic elements of the signal (sounds for speech, strokes for handwritten text) and *n*-grams language models (LM) are used to model word sequences (Jelinek, 1998). From this viewpoint, HTR systems fit in the Natural Language Processing paradigm. Therefore, many features that are usual to ASR systems (such as the use of training data for HMM and *n*-grams) are common to HTR systems as well.

The similarities between the two types of systems make possible to combine them easily into a multimodal system that may obtain a more reliable final hypothesis, since two different data sources (handwritten text and speech) can be used. In fact, previous attempts in combining handwritten input and speech input have been done (Liu and Soong, 2009), but most of them focus on the use of on-line handwritten text. For instance, in (Suhm et al., 2001; Liu and Soong, 2006), different speech-handwriting fusion methods are explored for (non-interactive) post-editing and for interactively correcting the output of a speech recogniser, respectively. In (Medjkoune et al., 2011) speech and on-line

* Corresponding author. Tel.: +34 96 387 70 07 73529; fax: +34 96 387 73 59.
E-mail addresses: valabau@iti.upv.es (V. Alabau), cmartine@dsic.upv.es (Carlos-D. Martínez-Hinarejos), vromero@iti.upv.es (V. Romero), alagarda@iti.upv.es (Antonio-L. Lagarda).

0167-8655/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.patrec.2012.11.007

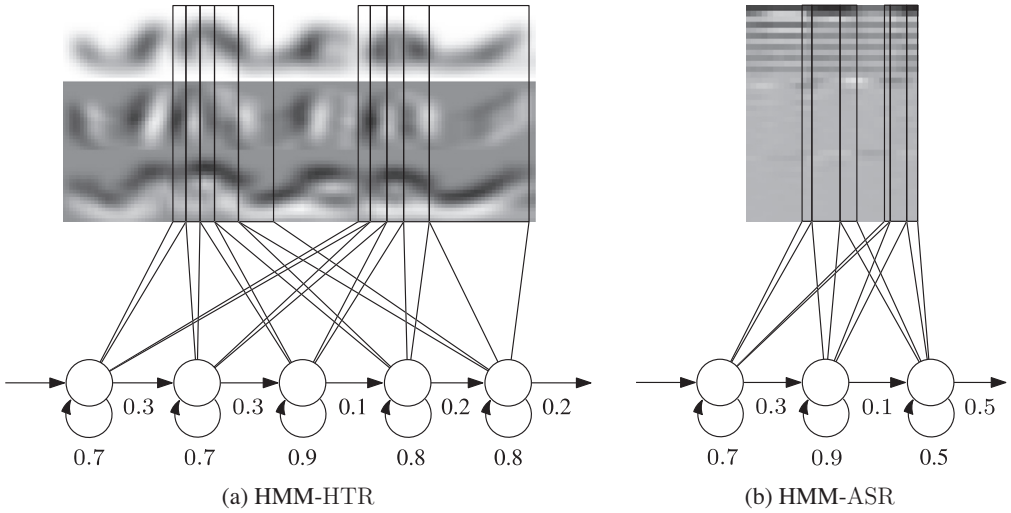

Fig. 1. Example of 5-states HMM for HTR (left) and 3-states HMM for ASR (right) modelling (sequences of feature vectors) instances of the character "a" and the phoneme "a", respectively, within the Spanish word "saca". The states are shared among all instances of characters/phonemes of the same class.

handwriting are combined for isolated mathematical symbol recognition with very encouraging results. Moreover, in (Humm et al., 2009) a fusion strategy based on the joint modelling of both streams is presented for user authentication. In addition, Pastor-i-Gadea and Paredes (2010) integrated the outputs from off-line and on-line HTR, but the best approach found was to use a simple naive-Bayes approach. More recently, the *HVR Grand Challenge Workshop* at the International Conference on Multimodal Interaction 2012[1] aims at finding new algorithms for helping speech recognition with handwritten gestures. However, the results were not available at the moment of writing this paper.

In a previous work (Alabau et al., 2011) a first attempt of combining off-line HTR and ASR systems showed promising results. The method consisted basically of restricting the ASR decoding process based on the results of the HTR decoding. In this work we extend the process described by Alabau et al. (2011) towards two different directions: we explore the effect of using the different modalities (HTR and ASR) as starting modality, and we study the iterative use of the process.

The paper is organised as follows: Section 2 describes the fundamentals of HTR and ASR systems, Section 3 explains the use of the HTR decoding to improve the ASR recognition, Section 4 summarises the experimental set-up, Section 5 shows the results, and Section 6 provides the main conclusions and future work lines in this field.

2. Systems overview

Several approaches have been proposed in the literature for HTR that resemble the noisy channel approach that is currently used in ASR. Consequently, HTR systems are based on hidden Markov models (HMM) (Marti and Bunke, 2001; Toselli et al., 2004; Plötz and Fink, 2009), and most recently, on recurrent neural network (RNN) (Graves et al., 2009) or hybrid systems using HMM and neural networks (HMM-RNN) (España-Boquera et al., 2011) with encouraging results (Grosicki and El-Abed, 2011). The HTR system used in this paper is based on HMM (Jelinek, 1998). This approach is the most consolidated and widely used; moreover, recent experiments carried out in (Romero et al., in press) showed that, depending on the task, HMM may produce better results than those obtained with RNN.

HMM are used here in the same way as they are used in the current ASR systems (Rabiner, 1989). The most important differences lay in the type of input sequences of feature vectors: while in the case of off-line HTR are line-image features, the input sequences for ASR represent acoustic data. Fig. 1 shows an example of how a HMM models two feature vector subsequences pertaining to the character "a" and the phoneme "a".

The problem of both handwriting and speech recognition can be formulated as the problem of finding the most likely word sequence, $\boldsymbol{w} = (w_1, w_2, \ldots, w_{|\boldsymbol{w}|})$, for a feature vector sequence $\boldsymbol{x} = (x_1, x_2, \ldots, x_{|\boldsymbol{x}|})$ describing a text image or speech signal along its corresponding horizontal or time axis i.e., $\boldsymbol{w} = \operatorname{argmax}_{\boldsymbol{w}} P(\boldsymbol{w}|\boldsymbol{x})$. Using the Bayes' rule we can decompose this probability into two probabilities, $P(\boldsymbol{x}|\boldsymbol{w})$ and $P(\boldsymbol{w})$, representing morphological or acoustical knowledge, and syntactic knowledge, respectively:

$$\hat{\boldsymbol{w}} = \operatorname*{argmax}_{\boldsymbol{w}} P(\boldsymbol{w}|\boldsymbol{x}) = \operatorname*{argmax}_{\boldsymbol{w}} P(\boldsymbol{x}|\boldsymbol{w}) P(\boldsymbol{w}) \tag{1}$$

$P(\boldsymbol{x}|\boldsymbol{w})$ is typically approximated by concatenated character/phoneme models, usually HMM, and $P(\boldsymbol{w})$ is approximated by a word LM, usually n-grams (Jelinek, 1998).

Each character/phoneme is modelled by a continuous density left-to-right HMM with a Gaussian mixture per state. This mixture serves as a probabilistic law to the emission of feature vectors on each model state. The optimum number of HMM states and Gaussian densities per state are tuned empirically. Each lexical word is usually modelled by a stochastic finite-state automaton (SFSA), which represents all possible concatenations of individual character/phonemes to compose the word. By embedding the character/phoneme HMM into the edges of this automaton, a lexical HMM is obtained. The model parameters can be easily trained from samples (handwritten text image or speech utterance) accompanied by the transcription of these samples into the corresponding sequence of characters/phonemes. This training process is carried out by using a well known instance of the EM algorithm called Forward–Backward or Baum–Welch. On the other hand, text lines or sentences are modelled using smoothed word n-grams, estimated from the training transcriptions of the text images.

Once all the character/phoneme, word and language models are available, recognition of new test sentences can be performed. Thanks to the homogeneous finite-state nature of all these models, they can be easily integrated into a single global model on which a search process is performed for decoding the input feature vectors sequence into an output word graph. This search is efficiently carried out by using the Viterbi algorithm.

The two implemented systems (HTR and ASR) are presented in detail in Subsections 2.1 and 2.2. Subsection 2.3 defines the

[1] http://speech.ddns.comp.nus.edu.sg/HVRGrandChallenge2012/.