A HMM-Based Approach to Recognize Ultra Low Resolution Anti-Aliased Words

Farshideh Einsele, Rolf Ingold, and Jean Hennebert

Université de Fribourg, Boulevard de Pérolles 90, 1700 Fribourg, Switzerland
{farshideh.einsele, rolf.ingold, jean.hennebert}@unifr.ch

Abstract. In this paper, we present a HMM based system that is used to recognize ultra low resolution text such as those frequently embedded in images available on the web. We propose a system that takes specifically the challenges of recognizing text in ultra low resolution images into account. In addition to this, we show in this paper that word models can be advantageously built connecting together sub-HMM-character models and inter-character state. Finally we report on the promising performance of the system using HMM topologies which have been improved to take into account the presupposed minimum length of each character.

1 Introduction

The explosive growth of the World Wide Web has resulted in a colossal data collection with billions of electronic documents. These documents often contain images with textual information providing very high semantic value which could be used for indexing. According to a study done in 2001, of the total number of words visible on a WWW page, 17% are in image form and 76% of these words do not appear elsewhere in the encoded text [2]. Furthermore, the `ALT` tag, which is recommended for describing an image in HTML language, is frequently not or wrongly used [7]. Web indexation engines would clearly benefit from having access to a so-called Web-image-OCR that could automatically recognize text embedded in images.

One approach is actually to use classical scanner-based OCR systems and to feed them with web images. However, these OCR systems are not trained nor tuned to treat such ultra low resolution images (< 100 dpi) with small point sizes (< 12 points) and with anti-aliasing artefacts. Furthermore, the detection of text area is more difficult in the case of web images as it is frequently surrounded or even superposed to other graphical objects. Acknowledging these difficulties, several related works have been proposing pre-processing methods to transform the image into a representation that is better suited for classical OCR systems [5] [1] [9]. Notwithstanding, most of these studies dealt with bi-level (black and white) images and the proposed methods did not address the specific variabilities of text embedded in web images.

In our approach, instead of focussing on pre-processing methods and relying on classical OCR systems, we propose to build a dedicated characters- and

A. Ghosh, R.K. De, and S.K. Pal (Eds.): PReMI 2007, LNCS 4815, pp. 511–518, 2007.

words-recognizer to handle the specificities of text embedded in web images, i.e. ultra low resolution rendered text images with small point sizes and anti-aliasing artefacts. In this paper, we propose to use a recognition system based on Hidden Markov Models (HMMs). HMMs are well known modelling tools widely used in the fields of handwriting or speech recognition, see for example [10]. HMMs have also been successfully used in classical OCR systems, with the advantage of being able to recognize connected characters such as arabic language [6]. However, they have never been proposed, to the best of our knowledge, for the recognition of text embedded in gray-level images with anti-aliasing. Furthermore, we believe that our approach differs from the classical use of HMMs for OCR tasks since we address the specificities of ultra low resolution rendered text images by applying the following procedures:

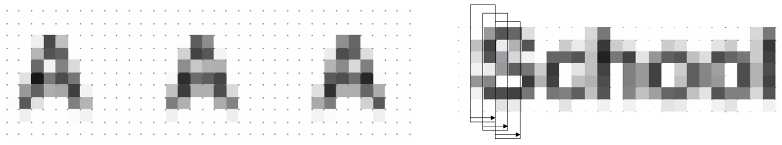
- We use a specific feature extraction module based on moments computation instead of the classical projection profiles used in OCR systems. This choice is guided to handle the limited amount of pixels that can be used as input.
- The probability density functions used in the HMMs are trained on ultra low-resolution characters to incorporate the specificities of the features such as the anti-aliasing and downsampling noise, i.e. we don't attempt to remove or reduce the anti-aliasing effect, instead, we model it explicitely.
- The topology chosen for the HMM will model inter-character regions. The associated models will also incorporate the specificities of these regions in terms of anti-aliasing noise due to the close proximity of adjacent characters.

In Section 2 of this paper we give a brief description of the specificities of text images encountered in web images. In section 3, we introduce the word recognizer system we have built and we describe the feature extraction as well as the different HMM topologies that have been used in this work. In section 4 we describe the task used to evaluate the performance of the recognizer and we then report on the results of 8 different configurations of the recognizer. Conclusions and plans for future work are finally presented.

2 Challenges for Web-Image-OCR

Character level. As shown on Fig. 1(a), a character embedded in a web image presents specific characteristics: (1) the character has an *ultra* low resolution, usually smaller than 100 dpi with small point sizes frequently between 6 and 12 points, (2) the character has artefacts due to anti-aliasing filters (3) the same characters can have multiple representations due to the position of the down sampling grid.

In our previous work [4], we have been evaluating a system aiming at the identification of such isolated characters. The objective was to assess the feasibility of recognizing such characters and to evaluate the difficulty of the task. We proposed to use a feature extraction based on first and second order central moments coupled with a statistical model based on multivariate Gaussian density functions. Encouraging results were obtained with overall recognition rates above 99.5%, consistently observed on 168 different fonts. While encouraging, these results do not correspond to a realistic situation as we did the assumption that characters were isolated, which is not the case in practice (see next section).



(a) Character level (b) Word level

Fig. 1. (a) Example of anti-aliased, down sampled character '`A`' with different grid alignments. (b) Low resolution version of word '`School`' and illustration of the sliding window used for the feature extraction.

Word level. Fig. 1(b) illustrates a more realistic example of characters composing a word in a low resolution image. As can be observed, there are no character interspaces available to segment characters within the word. Furthermore, the anti-aliasing artefact of adjacent characters is superposing. This contextual effect is clearly a new extra source of variability on the character samples.

Therefore, as an extension of our isolated character evaluation presented in the previous section, we wanted to measure the impact of this contextual variability [3]. We used the same classification system on the same task, assuming that character segmentation was known. The only difference was coming from the contextual noise as an extra source of variability. We observed an overall character recognition rate of 98.5% which is less than the former result but still encouraging.

However, the system used to obtain these results is not realistic as it intrinsically implies that the segmentation of characters inside of the word is known. A direct observation of the word "School" in Fig. 1(b) can convince us that well-known pre-segmentation methods used in classical OCR systems can not be applied anymore in our case [8]. Acknowledging this fact, we opted to build a new recognizer system that is able to solve the classification problem simultaneously with the segmentation problem.

3 System Description

In a similar way as what is frequently done in or cursive handwriting or speech recognition modeling, we decided to build our Web-OCR system using a sliding-window feature extraction feeding an HMM based classifier. This combination has the clear advantage that the segmentation and recognition problem can be solved simultaneously. HMMs also bring further advantages by allowing to naturally include linguistic knowledge such as lexical or grammatical models inside of the framework.

Before going to the more complex task of connected word recognition involving more advanced linguistic models, we decided to build and evaluate a HMM system designed for the task of isolated word image recognition. Such a task actually makes sense as we can reasonably assume that words can be isolated with classical image processing techniques, even though they are low resolution and have small font sizes. The details of our system are described hereafter.

3.1 Feature Extraction

HMMs are basically modelling ordered sequences of features that are function of a single independent variable. Inspired by feature extraction used for speech or cursive handwriting recognition, we decided to compute a naturally left-right ordered sequence of features by sliding an analysis window on top of the word. Therefore the independent variable is simply, in our case, the x-axis. As illustrated on Fig. 1(b), we used a 2 pixels length window shifted 1 pixel right. In each analysis window, a feature vector of 8 components including the first and second order central moments is computed. The choice of moments as features was directed by their good discriminative representation as observed in our previous studies (see section 2). The word image is then represented by a sequence of feature vector $X = \{x_1, x_2, ..., x_N\}$ where x_n is a 8 component vector.

3.2 Hidden Markov Models

The HMM will model the likelihood of the observation sequence X given the model parameters associated to a word i. By applying the usual simplifying assumption of HMMs (see for example [10]), the likelihood $P(X|M_i)$ of X given the model M_i can be expressed as the sum, over all possible paths of length N, of the product of emission probabilities and transition probabilities measured along the paths. Alternatively, the Viterbi criterion can also be used, stating that instead of considering all potential paths through the HMM, only the best path is taken into account, i.e. the path that maximizes the product of emission and transition probabilities. In this work, we have based our training and testing approaches using the Viterbi criterion.

For sake of flexibility, we have chosen to associate HMM states to characters. Doing this, any word can be modelled by an HMM where the corresponding states are simply connected together.

Fig. 2 illustrates the single word HMM-recognizer that has been used to perform our evaluation. We have built a vocabulary of 520 words that were different from the words used in the training set. The words were picked from a dictionary in a way to represent all characters according to their natural occurrence. Each test image has been synthetically generated using the *Verdana* font. From a high resolution version, the image was then downsampled to obtain the equivalent of 9 point size with a resolution of 72 dpi, applying anti-aliasing filters and different grid alignments, such as illustrated on Fig. 1(b). The recognizer first computes the feature vectors that are fed to the 520 competing HMMs. The computation of the likelihood for each model is performed using the Viterbi criterion. As, in our test, all words are equally a priori probable, the winning model is simply the one that leads to the maximum likelihood value.

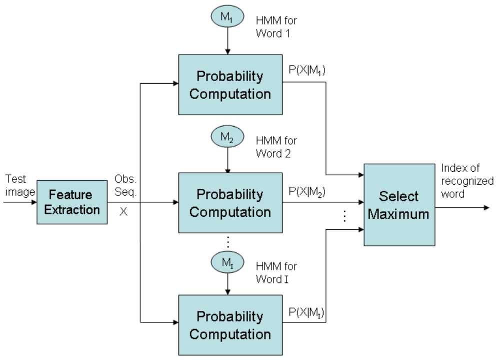


Fig. 2. Single word HMM-Recognizer

4 Topologies

We investigated different topologies for the HMMs.

4.1 Simple Left-Right

This topology was chosen for its simplicity and was, for us, the first configuration to investigate. Fig. 3, illustrates this topology (top). As the performances

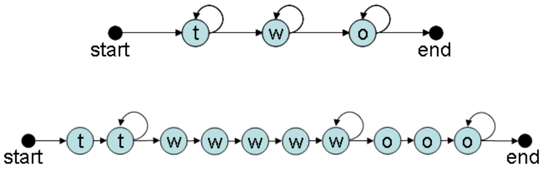


Fig. 3. Simple left-right topology (top) and left-right minimum duration tolopogy (bottom)

obtained with this topology were not so convincing, we inspected some state sequences obtained on mis-recognized words. An example of state sequence for word 'two' calculated with simple left-right topology is reproduced below:

- Genuine word= two
- Recognized word = one
- Best state sequence for word one: o o o o o o o o o o o n n e e e e
- Best state sequence for word two: t t t t w w w w o o o o o o o o o
- Ground truth state sequence for word two: t t t t w w w w w w w o o o o o

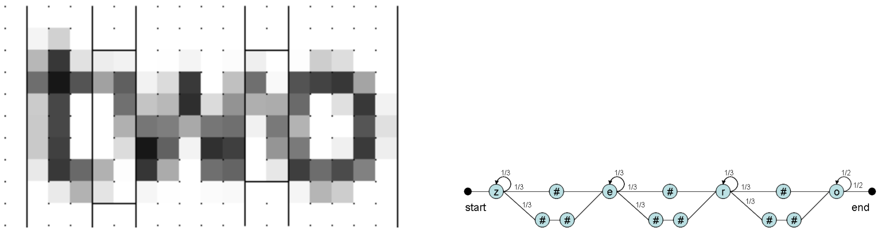
What is happening can be explained as follows. By nature, the HMM is trying to maximize its likelihood. It will then associate few observations to states that gives low emission probabilities while it will spend more observations in states giving high emission probabilities. The state sequence for the winning word 'one' is clearly spending only two observations in character model 'n'. Character 'n' is of course longer than 2 pixels (actually 3 pixels considering the width of the analysis window) in our images.

4.2 Left-Right Minimal-Duration

In order to avoid phenomena as the one described above, one can introduce so-called minimum duration topologies. This topology is simply obtained by repeating a state a number of time obliging the Viterbi algorithm to spend a minimal amount of observations in the same category of character while the last character state has the possibility to be repeated for an unlimited amount of time. Fig. 3 illustrates such a topology for word 'two' (bottom). In our configuration, the minimum duration values have been obtained from the a priori known font metric information. We have to underline that such minimum duration values are dependent to a given font, leading to a system tuned for a specific font.

4.3 The Inter-character Model

According to our previous study of characters in context (see section 2 and [3]), the left and right borders of characters are influenced by the left and right borders of their adjacent characters. We have observed that this *noisy* zone between two adjacent characters includes up to three pixels for font sizes between 6 to 12 points. We have decided to treat this anti-aliasing noise as an additional character model and perfomed training as for another character. We therefore called this new pseudo character model the *inter-character* ('#' in our figures). Fig. 4(a) illustrates the word 'two' and the corresponding inter-character zones.



(a) (b)

Fig. 4. (a) Inter-character zones from word 'two' (b) Inter-Character topology

The previous two HMM-topologies have been modified to take into account the 'inter-character' model. Fig. 4(b) illustrates this modification for the simple left-right HMM-topology.

5 Evaluation Tests

According to the topologies introduced above, different evaluation tests have been performed. We also experimented with two implementations of the Gaussian probability density functions used to estimate the emission probabilities. The first one is using the regular *full* covariance matrix. The second one is using a simplified *diagonal* covariance matrix, making the extra assumption that the components of the feature vectors are de-correlated. While this assumption is potentially too restrictive, it allows a much faster computation of the emission probabilities.

Table1 illustrates the results of theses evaluation tests. The table is divided into two parts. The first part presents results obtained with our previously described two HMM-topologies. The second part presents results obtained from the same HMM-topologies, which have been extended with the inter-character model. Our main observations are the following:

- As expected, we see clearly an improvement coming from the introduction of minimum duration topologies.
- Using full covariance matrices leads to better results than with diagonal covariance matrices.
- The most significant improvement is coming from the introduction of the inter-character model into the HMM topologies.

These encouraging results indicate clearly that HMMs are strongly suitable to simultaneously segment and recognize ultra low resolution words, provided that the right topologies and models are used.

We have to notice that the results are obtained for a test corpus of 520 words. The performance for the HMM using the inter-character model are potentially not anymore significantly different. We plan to use a larger number of test samples and a more difficult task, increasing the test corpus to several thousand words, to better put in evidence system differences.

Table 1. Recognition rates , **without inter-char** (with inter-char)

	diag. covariance	full covariance
Simple left-right	**52%** (> 99%)	**56%** (> 99%)
Minimum duration	**60%** (> 99%)	**79%** (> 99%)

6 Conclusion and Future Work

Most of the approaches to solve the problem of recognizing text embedded in web images are based on pre-processing the images in order to feed them in classical OCR systems. We have presented in this paper a competing approach that is not based on pre-processing of the images but that aims at directly model the specificities of these images: gray-level, ultra low resolution and anti-aliased. An extra piece of difficulty to the task is coming from the fact that adjacent characters cannot be anymore pre-segmented using well-known image processing algorithms. We therefore have proposed to base our system on HMMs that ally powerful statistical models with the ability of performing segmentation in the same time as recognition. While we acknowledge that our experiments are currently performed on synthetic data instead of real-world data, the evaluation tests that we performed with the HMM system brings promising results and clearly show that HMMs are able to simultaneously segment and recognize ultra low resolution words.

In our future works, we plan to improve the HMM system using more powerful emission probability estimators such as multi-Gaussian models. Such models would allow us to relax the assumption that the features are distributed according to a unique Gaussian. In addition to this, we plan to use multi-state character models that will allow us to capture more finely the characteristics of multi-stroke characters. Finally, we plan to move towards more complex recognition tasks involving much larger vocabularies or even open vocabularies, using multi-font context and using data extracted from real-world web images.

References

1. Antonacopoulos, A., Karatzas, D.: Text extraction from web images based on a split-and-merge segmentation method using color perception. In: Proc. of ICPR 2004, Cambridge, UK (August 2004)
2. Antonacopoulos, A., Karatzas, D., Lopetz, J.O.: Accessing textual information embedded in internet images. In: Proc. of Electronic Imaging II, San Jose, California, USA (January 2001)
3. Einsele, F., Hennebert, J., Ingold, R.: Towards identification of very low resolution, anti-aliased characters. In: Proc. of ISSPA 2007, Sharjah, UAE (February 2007)
4. Einsele, F., Ingold, R.: A study of the variability of very low resolution characters and the feasibility of their discrimination using geometrical features. In: Proc. of 4th Enformatika Int. Conf. on Pattern Recognition and Computer Vision, Istanbul, Turkey, pp. 213–217 (June 2005)
5. Lopresti, D., Zhou, J.: Locating and recognizing text in www images. Information Retrieval 2(2/3), 177–206 (2000)
6. Lu, Z., Bazzi, I., Kornai, A., Makhoul, J., Natarajan, P., Schwartz, R.: A robust, language-independent ocr system. In: Proc. 27th IAPR Workshop, vol. 3584, pp. 96–104 (January 1999)
7. Munson, E.V., Tsymbalenko, Y.: Using html metadata to find relevant images on the web. In: Web Document Analysis
8. Nagy, G.: Twenty years of document image analysis in pami. IEEE Transactions on Pattern Analysis and Machine Intelligence 22, 38–62 (2000)
9. Perantonis, S.J., Gatos, B., Maragos, V.: A novel web image processing algorithm for text area identification that helps commercial ocr engines to improve their web recognition accuracy. In: Proc. of the second Int. Workshop on Web Document Analysis, Edinburgh, United Kingdom (August 2003)
10. Rabiner, L., Juang, B.-H.: Fundamentals Of Speech Recognition. Prentice Hall, Englewood Cliffs (1993)