A Bi-modal Handwritten Text Corpus: baseline results

Moisés Pastor-i-Gadea, Alejandro H. Toselli, Francico Casacuberta, Enrique Vidal
Universitat Politécnica de Valencia, Camino de Vera s/n
{*moises,ahector,fcn,evidal*}*@iti.upv.es*

***Abstract*—Handwritten text is generally captured through two main modalities: *off-line* and *on-line*. Smart approaches to handwritten text recognition (HTR) may take advantage of both modalities if they are available. This is for instance the case in computer-assisted transcription of text images, where on-line text can be used to interactively correct errors made by a main off-line HTR system. We present here baseline results on the biMod-IAM-PRHLT corpus, which was recently compiled for experimentation with techniques aimed at solving the proposed multi-modal HTR problem, and is being used in one of the official ICPR-2010 contests.**

I. Introduction

Handwritten text is one of the most natural communication channels currently available to most human beings. Moreover, huge amounts of historical handwritten information exist in the form of (paper) manuscript documents or digital images of these documents.

When considering handwritten text communication nowadays, two main modalities can be used: *off-line* and *on-line*. Several authors have studied the possibility of using *both* the on-line trajectory *and* a corresponding off-line version of this trajectory (its *"e-ink"*) [1], [2], [3]. This *multi-modal* recognition process has been reported to yield some accuracy improvements over using only the original on-line data.

An interesting scenario where a related, but more challenging bi-modal (on/off-line) fusion problem arises is Multi-modal Computer Assisted Transcription of Text Images, called "MM-CATTI" in [4]. In this scenario, errors made by an off-line HTR system are immediately fixed by the user by means of on-line pen strokes or text written on a tablet or touch-screen [4]. Clearly, most of these corrective pen-strokes are in fact on-line text aimed to fix corresponding off-line words that have been miss-recognized by the off-line HTR system. This allows taking advantage of both modalities to improve the feedback decoding accuracy and the overall multi-modal interaction performance.

In order to ease the development of adequate techniques for such a challenging bi-modal HTR recognition task, the *"biMod-IAM-PRHLT"* corpus [5] was recently compiled.

In this work we briefly overview the *biMod-IAM-PRHLT* corpus (Section III) and available techniques possibly useful to approach the proposed multi-modal stream recognition problem (Section II). Then, both uni-modal and bi-modal baseline results on this corpus are presented (Section IV). These results revise and extend those presented in the technical report accompanying the *biMod-IAM-PRHLT*corpus [5]. Finally, some conclusions are presented (Section V).

Of course, while MM-CATTI is our main motivation for this study many other applications exist with similar underlying multi-modal problems.

II. Multi-modal (stream) recognition

For simplicity, we will assume a classification scenario into C classes (words in isolated-word handwriting recognition) and two data streams x and y (sequences of feature vectors representing an on-line pen tip trajectory and an off-line text image – see section IV). Both streams are assumed to correspond to a unique sequence of events (e.g, a sequence of characters constituting a word).

In a uni-modal scenario (for example, using x only), the minimum error *Bayes classification rule* can be formulated as the search for a c that maximizes $\Pr(x|c) \cdot \Pr(c)$. For the kind of data here considered, the well known *hidden Markov models* (HMMs) have been successfully used to implement $\Pr(x|c)$ [6], [7], [8].

In a general classification scenario, an independent HMM is associated to each class c. However, in many problems, such as speech recognition [9], [10] and handwriting text recognition [6], [7], [8], the HMM for each class (word) is built from HMMs of word constituents. If a word c is composed of K characters and if we assume, for simplicity, that a segmentation of x into K segments is given: $x = \bar{x}_1, \ldots, \bar{x}_K$, then, we can write:

$$P(x|c) \quad = \quad \prod_{k=1}^{K} P(\bar{x}_k|c_k) \tag{1}$$

where each segment $\bar{x}_k$ is generated by a character HMM c_k according to $P(\bar{x}_k|c_k)$.

With respect to using independent whole word models, this approach has the advantage that the total number of different model parameters is generally much smaller, leading to better estimated probabilities.

In the bi-modal scenario, the decision rule entails the search for a c that maximize $\Pr(x, y|c) \cdot \Pr(c)$. Different approaches have been proposed so far depending on the assumptions made on $\Pr(x, y|c)$. The *naive Bayes classifier* is perhaps the simplest scheme for multi-modal recognition of two data streams (x, y). It is based on strong independence assumptions on $\Pr(x, y|c)$, leading to $\Pr(x, y|c) = \Pr(x|c) \cdot \Pr(y|c)$ [11].

1051-4651/10 $26.00 © 2010 IEEE
DOI 10.1109/ICPR.2010.476

Of course, also in this case, when x and y represent event sequences (words), HMMs are good choices to model $\Pr(x|c)$ and $\Pr(y|c)$. Moreover, the naive Bayes approach can also be applied now at the sub-word (character) level [9]. If a word c is composed of K characters and assuming for simplicity that segmentations of x and y into K segments are given ($x = \bar{x}_1, \ldots, \bar{x}_K, y = \bar{y}_1, \ldots, \bar{y}_K$), $\Pr(x, y|c)$ can be approached as:

$$P(x, y|c) = \prod_{k=1}^{K} P(\bar{x}_k|c_k) \cdot P(\bar{y}_k|c_k) \tag{2}$$

where $P(\bar{x}_k|c_k)$ and $P(\bar{y}_k|c_k)$ are given by on- and off-line HMMs of the k-th character in c for the first and the second modalities, respectively.

In practice, the models of each modality are considered in parallel but they share the first and the last state. The decoding is carried out independently in each model using the corresponding data stream asynchronously, but the process is synchronized in the shared states. Note that, since the segmentation into K segments is not known in advance, this entails a difficult search problem that can not be directly solved by the Viterbi algorithm [12].

Many other models and techniques have been proposed for multi-modal stream fusion. See [13], [9], [10] among others. These sophisticated approaches aim to model the interdependencies which are neglected by the naive Bayes assumption. However, in this paper only results using the naive Bayes classifier, both at the word and the character level, will be presented.

III. The biMod-IAM-PRHLT Corpus

The *"biMod-IAM-PRHLT"* corpus [5], has been recently compiled as a test-bed for the above outlined multi-modal decoding approaches. The corpus is simple, while still entailing the essential challenges of the bi-modal HTR problems discussed in Section I.

The samples of the biMod-IAM-PRHLT corpus are word-size segments from the publicly available IAMDB off-line and on-line corpora [14], [15]. The off-line word images were semiautomatically segmented at the IAM (FKI) from the original page- and line-level images. On the other hand, the on-line IAMDB was only available at the line-segment level. Therefore researchers of the PRHLT group segmented and extracted the adequate word-level on-line samples, as discussed in [5].

Figure 1 shows some examples of on/off-line word pairs contained in this corpus.

Figure 1. Four examples of on-line/off-line pairs of words from the biMod-IAM-PRHLT corpus.

The corpus is explicitly partitioned into training and test sub-corpora. In addition to the test data included in the current version (referred to as *test1*), more on-line and off-line test samples were produced, but they are currently held-out to be used for benchmarks such as the ICPR-2010 Bi-modal HTR contest.

Main figures of the publicly available part of the *biMod-IAM-PRHLT* corpus are shown in Table I (See [5] for more information).

Table I
Basic statistics of the biMod-IAM-PRHLT corpus and their standard partitions.

	on-line	off-line
Word classes (vocabulary)	519	519
training samples	8 342	14 409
test1 samples	519	519
total samples	8 861	14 928

IV. Baseline experiments

Experiments were carried out to establish baseline accuracy figures for the biMod-IAM-PRHLT corpus. In these experiments, fairly standard preprocessing and feature extraction procedures and character-based HMM word models were used.

In the on-line case, each trajectory is processed through only three simple steps: pen-up points elimination, repeated points elimination, and noise reduction (by simple low pass filtering). Each preprocessed trajectory is transformed into a new temporal sequence of 6-dimensional real-valued feature vectors [16]. These features are: normalized *vertical position*, normalized first and second time *derivatives* and *curvature*.

On the other hand, off-line image preprocessing consists on median filter noise removal, slope and slant corrections and size normalization. To transform the off-line signal into a sequence of feature vectors, a grid of $N \times M$ squared cells is applied to each text image ($N = 20$ in this work). For each cell, three features are calculated: normalized *gray level*, and horizontal and vertical *gray level derivatives*. The computation of these features is smoothed by convolution with a 2-d 5×5 cells Gaussian filter. At the end of this process, a sequence of M 60-dimensional vectors is obtained [17].

The HTR system used both for on- and off-line decoding is based on continuous density HMMs, trained using the well-known HTK toolkit [18]. The number of states of each character HMM (N_{s_c}) is estimated depending on the average character segment length as:

$$N_{s_c} = \bar{S} \frac{\bar{n}_{f_c}}{\sum_c \bar{n}_{f_c}} \tag{3}$$

where $\bar{S}$ is a given average number of states across all models and $\bar{n}_{f_c}$ is the average number of feature vectors per character. Word models are concatenations of character models as discussed in Section II.

With these on- and off-line models, two types of classification experiments were carried out: *unimodal* using individual on- and off-line HMMs and *bi-modal*, based on the naive Bayes approach. According to the definition of the biMod-IAM-PRHLT corpus, $\Pr(c)$ is assumed to be uniform (equiprobable words).

A. Results

Classification results for *test1*, using individual on-line and off-line HMMs are presented in Figure 2. The lowest on-line and off-line error rates achieved were 6.6% and 27.6%, using 12 average states per model and 8 Gaussians per state, and 8 average states per model and 64 Gaussians per state, respectively.

If we consider each HMM as a whole model (i.e. without using character-level scores), we can compute some error bounds. The simplest one corresponds to an *oracle* which, for each test sample, selects a modality for which the classification result is correct, if any, and a random modality otherwise. For *test1*, such a *best-modality oracle* error rate was 2.3% .

To obtain baseline bi-modal results, we used the naive Bayes classifier and the above discussed on-line and off-line HMMs. In practice, we used a weighted-log version of this classifier (also assuming uniform priors), which aims at balancing the relative reliability of the on-line (x) and off-line (y) models:

$$\hat{c} = \underset{1 \le c \le C}{\operatorname{argmax}} \left((1-\alpha) \cdot \log P(x|c) + \alpha \cdot \log P(y|c) \right) \qquad (4)$$

When this weighted scheme is adopted, a more elaborate oracle can be defined which corresponds to a true lower-bound of the error-rate for classifiers based on whole-word (weighted) scores (*"word-level lower bound"*): for a sample (x, y), with true class label r, an error is produced only if $\exists c \neq r : \log P(x \mid r) < \log P(x \mid c)$ *and* $\log P(y \mid r) < \log P(y \mid c)$; otherwise an α may exist such that $\hat{c} = r$ in (4).

In these experiments, all the log-probability values (scores) were previously shifted so that both modalities have a fixed, identical maximum score, U. Then, to reduce the impact of low, noisy probabilities, all the scores lower than a given threshold T were truncated to that threshold. The chosen values for U and T are not important, but the difference $U - T$ is. In our case good results were obtained for $U - T = 1\,000$.

Results with character level naive-Bayes combination were also obtained. To this end, partial scores associated to each character HMM were used and specific character weights were determined for a few characters ("R", "j" and "x") whose weighting have proved to improve the results. Optimal weights for these characters (0.3, 0.2, and 0.1, respectively) were obtained using the available validation data (*test1*), so we must wait for definitive testing on the currently held-out test data (*test2*) to draw conclusions on this approach.

Figure 3 plots the results on *test1* for varying α. The best word-level error rate is 4.0%, which is higher than the lower bounds, but is significantly better (39% relative) than the 6.6% error rate achieved by on-line models alone. Using character-level weights a slightly lower error rate (3.5%) is obtained. The optimal accuracy is achieved for an off-line weight (0.09) which is much lower than the on-line wight (0.91). This is consistent with he relative accuracies of the individual modalities.

Finally in Table II summarizes the main results.

V. Conclusions

There are many pattern recognition problems where different streams represent the same sequence of events. This multi-modal representation offers a good opportunities to

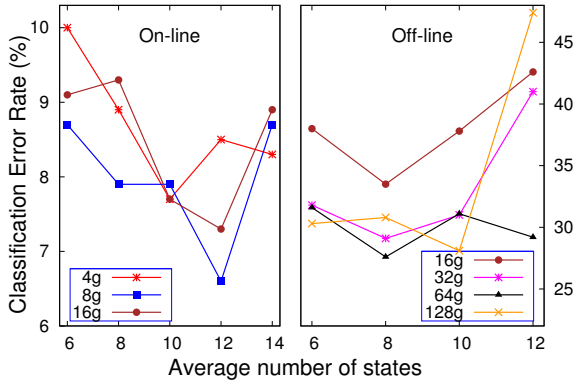

Figure 2. Classification Error on the *test1* partition as a function of the average number of states ($\bar{S}$), and the number of Gaussians per state.

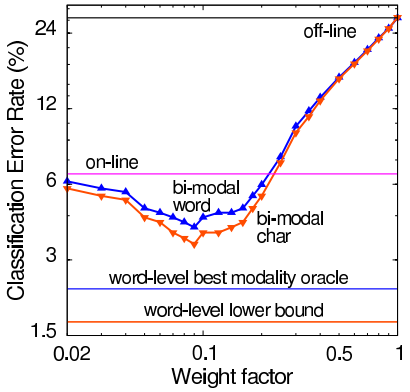

Figure 3. Bi-modal *test1* results (classification error rate %) using the word-level and character level naive Bayes classifiers as a function of the weight factor α.

Table II
BEST RESULTS (CLASSIFICATION ERROR RATE %) ON *test1* USING ON-LINE AND OFF-LINE CLASSIFIERS ALONE, THE THE WORD- AND CHARACTER-LEVEL NAIVE BAYES CLASSIFIERS AND BI-MODAL WORD-LEVEL ORACLES. THE RELATIVE IMPROVEMENT OVER THE ON-LINE-ONLY ACCURACY IS ALSO REPORTED.

Classifier	Error (%)	Rel. Improv.
off-line	27.6	–
on-line	6.6	–
word bi-modal	**4.0**	39%
character bi-modal	**3.5**	47%
word best mod. oracle	2.3	–
word lower bound	1.7	–

exploit the best characteristics of each modality to get improved classification rates. Recently, a controlled bi-modal corpus of isolated handwriting words was introduced in order to ease the experimentation with different models that deal with multi-modality. Here, baseline results on this corpus are reported that include uni-modal results, bi-modal results under a naive Bayes framework at the word-level and with a simple character weighting scheme. Oracle-based error lower bounds are also reported. From these results, we can conclude that multi-modal classification can help to improve the results obtained from the best uni-modal classification. In future works, more sophisticated techniques (Section II) will be applied to this corpus.

ACKNOWLEDGMENTS

Work supported by the Spanish MCIN grant TIN2009-14633-C03-01, the research programme Consolider Ingenio 2010: MIPRCV (CSD2007-00018) and by the Generalitat Valenciana under grant Prometeo/2009/014 .

REFERENCES

[1] A. Vinciarelli and M. Perrone, "Combining online and offline handwriting recognition," in *International Conference on Document Analysis and Recognition*, 2003, p. 844.

[2] M. Liwicki and H. Bunke, "Combining on-line and off-line bidirectional long short-term memory networks for handwritten text line recognition," in *Proceedings of the 11th Int. Conference on Frontiers in Handwriting Recognition*, 2008, pp. 31–36.

[3] C. Viard-Gaudin, P. Lallican, S. Knerr, and P. Binter, "The ireste on/off (ironoff) dual handwriting database," in *International Conference on Document Analysis and Recognition*, 1999, pp. 455 – 458.

[4] A. Toselli, V. Romero, M. Pastor, and E. Vidal, "Multimodal interactive transcription of text images," *Patter Recognition*, 2010.

[5] M. Pastor, E. Vidal, and F. Casacuberta, "A bi-modal handwritten text corpus," Instituto Tecnológico de Informática, http://prhlt.iti.es, Tech. Rep., Sep. 2009.

[6] I. Bazzi, R. Schwartz, and J. Makhoul, "An Omnifont Open-Vocabulary OCR System for English and Arabic," *IEEE Trans. on PAMI*, vol. 21, no. 6, pp. 495–504, 1999.

[7] A. Vinciarelli, S. Bengio, and H. Bunke, "Offline Recognition of Unconstrained Handwritten Texts Using HMMs and Statistical Language Models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 6, pp. 709–720, june 2004.

[8] A. H. Toselli, A. Juan, J. González, I. Salvador, E. Vidal, F. Casacuberta, D. Keysers, and 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.

[9] H. Bourlard and S. Duponet, "Subband-based speech recognition," in *Proceedings of the International Conference on Acoustic, Speech and Signal Processing*, April 21-24 1997, pp. 545–548.

[10] G. Potamianos, C. Neti, J. Luettin, and I. Matthews, *Issues in Visual and Audio-Visual Speech Processing*. MIT Press, 2004, ch. Audio-Visual Automatic Speech Recognition: An Overview.

[11] P. Verlinde, P. Druyts, G. Chollet, and M. Acheroy, "Applying bayes based classifiers for decision fusion in a multi-modal identity verification system," in *International Symposium "In Memoriam Pierre Devijver"*, February 1998.

[12] Y. Kessentini, T. Paquet, and A. B. Hamadou, "Off-line handwritten word recognition using multi-stream hidden markov models," *Pattern Recognition Letters*, vol. 31, no. 1, pp. 60–70, 2010.

[13] S. Bengio, "Multimodal speech processing using asynchronous hidden markov models," *Information Fusion*, vol. 5, pp. 81–89, 2004.

[14] U. Marti and H. Bunke, "A full english sentence database for off-line handwriting recognition." in *In Proc. of the 5th Int. Conf. on Document Analysis and Recognition*, 1999, pp. 705–708.

[15] M. Liwicki and H. Bunke, "Iam-ondb - an on-line english sentence database acquired from handwritten text on a whiteboard." in *8th Intl. Conf. on Document Analysis and Recognition*, vol. 2, 2005, pp. 956–961.

[16] A. H. Toselli, M. Pastor, and E. Vidal, "On-Line Handwriting Recognition System for Tamil Handwritten Characters," in *3rd Iberian Conference on Pattern Recognition and Image Analysis*, ser. Lecture Notes in Computer Science. Girona (Spain): Springer-Verlag, June 2007, vol. 4477, pp. 370–377.

[17] A. H. Toselli, A. Juan, and E. Vidal, "Spontaneous Handwriting Recognition and Classification," in *Proceedings of the 17th International Conference on Pattern Recognition*, vol. 1, Cambridge, United Kingdom, Aug. 2004, pp. 433–436.

[18] S. Young, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book: Hidden Markov Models Toolkit V2.1*, Cambridge Research Laboratory Ltd, Mar. 1997.