Recognising the Chilean Month Word Database

Luiz S. Oliveira and Marisa Morita

Federal University of Parana (UFPR)
Department of Informatics (DInf)
R. Rua Cel. Francisco H. dos Santos, 100, Curitiba, PR, Brazil
lesoliveira@inf.ufpr.br

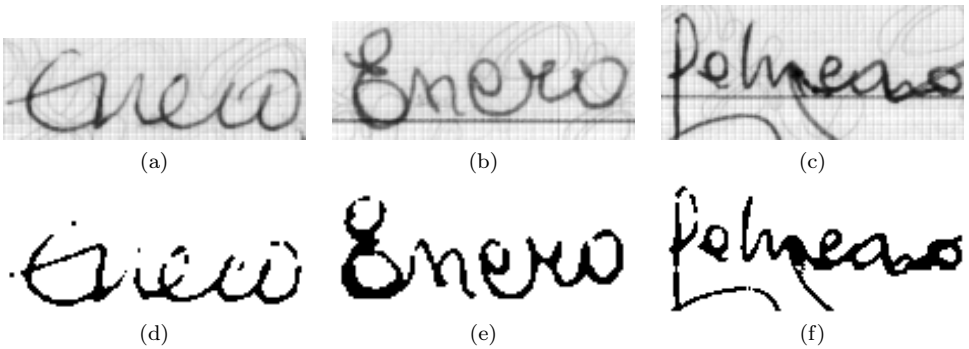
1 Introduction

In this report we describe the results of the experiments performed on the Chilean month word database. It is worth of remark that this dataset still is under construction, therefore the number of samples still is limited and not all the classes are available. To overcome these limitations, we have performed the recognition using the models trained with the Brazilian month words. Before presenting the recognition results, we present the preprocessing steps performed to remove baselines, background, and some noise.

2 Pre-processing

The pre-processing used so far is composed of three basic steps. First we convert the 256-graylevel image to a binary image using Otsu algorithm. We have tried out some other methods but Otsu seems to produce the best results. The second step is devoted to remove the baseline, which usually is located in the bottom but also may appear anywhere in the image. Some different strategies were tried out in this case, such as the morphological opening, but the by searching and removing for very long and thin components seems to be a good alternative. The final step intends to remove salt-and-pepper noise resulting from the binarization processes. This is done through a median filter of size 3. Figure 1 shows some examples of the database before and after the pre-processing. One possible solution for this kind of problem could be a strategy based on adaptive thresholding.

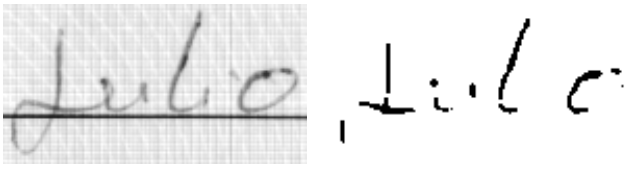
Im most cases the pre-processing produces good results. In some cases, though, specially when the pressure on the handwriting is weak, some parts of the image are lost. This is exemplified in Figure 2. In this case, it can be noticed that the baseline stroke is quite stronger than the handwriting itself. This has an important impact in the binarization process.



(a) (b) (c)

(d) (e) (f)

Figure 1: Some examples extracted from the database and their corresponding binary images.



(a) (b)

Figure 2: Some examples extracted from the database and their corresponding binary images.

3 Recognition Results

Based on our previous results (reported in Report #3), the best results using global features were achieved by the parallel model (Figure 3), in which the 12 HMM word models are built by concatenating all the corresponding characters. For this reason we decided to employ this model in this first experiment with the Chilean month word database.

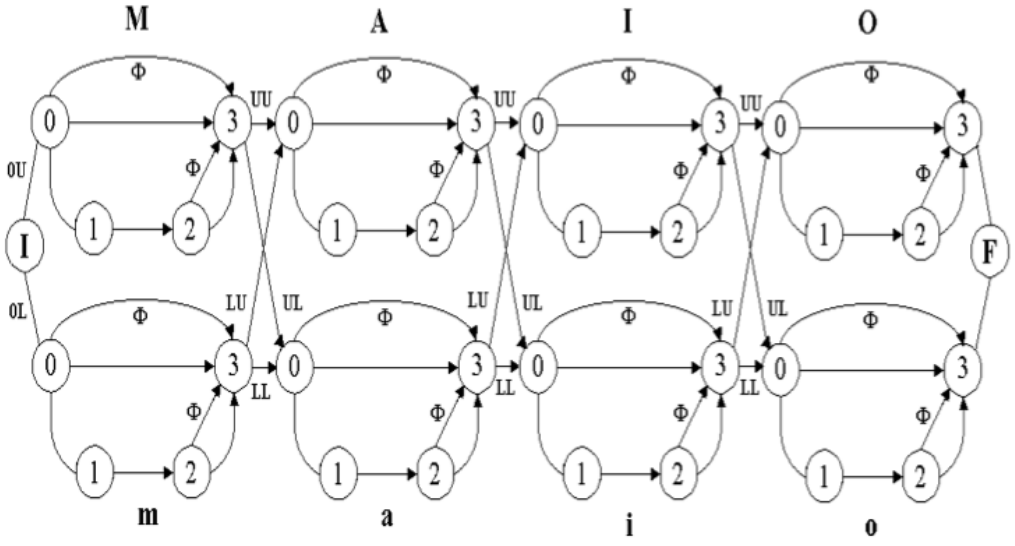
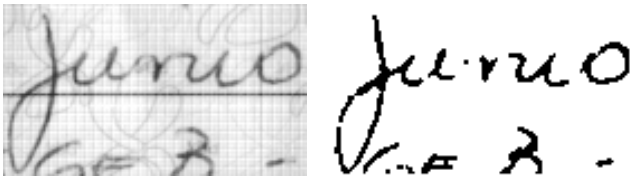


Figure 3: Training model of class "Maio"

Independently of the architecture choice, the first challenge to overcome is the lack of some

characters in the training set. As stated before, as we do not have enough data, we have used the Brazilian database to train the models. However, the Brazilian database does not have the characters “y” (appearing in Mayo) and “p” (appearing in Septiembre). in the 12 months. Since we are relying on Global features, we tried to find similar characters to replace those missing. To replace “y” and “p”, we have used “j” and “g”, respectively. In both cases, we have selected characters with descenders. This may work for cursive characters, but for upper-case this replacement is not straightforward.

Other aspect that compromises the performance on this database is the presence of spurious segments coming from other parts of the bank cheques. This kind of problem, which is illustrated in Figure 4, is a by-product of the segmentation and may have an impact in the feature extraction, hence, in the classification. A more elaborated pre-processing or segmentation strategy would be necessary to deal with this kind of problem.



(a) (b)

Figure 4: Segmentation problems.

We have performed the recognition of about 800 words available in the Chilean month word database. The recognition results for TOP1 and TOP3 are 42.2% and 74.7%, respectively. The confusion matrices for TOP1 and TOP3 are presented in Tables 1 and 2, respectively.

Table 1: Confusion Matrix (%) - TOP1

Class	1	2	3	4	5	6	7	8	9
1	**42.9**	19.8	3.3	4.4	2.2	18.7	6.6	2.2	0.0
2	14.9	**57.4**	7.4	4.3	0.0	5.3	5.3	2.1	3.2
3	32.6	21.1	**22.1**	1.1	6.3	10.5	1.1	5.3	0.0
4	38.9	14.7	4.2	**27.4**	1.1	3.2	5.3	4.2	1.1
5	22.9	6.3	22.9	1.0	**35.4**	8.3	2.1	1.0	0.0
6	14.4	12.4	2.1	1.0	1.0	**60.8**	8.2	0.0	0.0
7	19.0	9.0	1.0	1.0	1.0	12.0	**57.0**	0.0	0.0
8	17.2	21.2	4.0	10.1	3.0	3.0	5.1	**31.3**	5.1
9	0.0	11.8	5.9	11.8	0.0	0.0	0.0	11.8	**58.8**

Table 2: Confusion Matrix (%) - TOP3

Class	1	2	3	4	5	6	7	8	9
1	**68.1**	14.3	3.3	3.3	1.1	5.5	3.3	1.1	0.0
2	12.8	**76.6**	3.2	2.1	0.0	2.1	1.1	1.1	1.1
3	12.6	8.4	**66.3**	0.0	2.1	9.5	1.1	0.0	0.0
4	22.1	4.2	4.2	**61.1**	0.0	3.2	3.2	1.1	1.1
5	15.6	3.1	2.1	1.0	**70.8**	5.2	1.0	1.0	0.0
6	7.2	3.1	1.0	1.0	1.0	**86.6**	0.0	0.0	0.0
7	3.0	4.0	1.0	1.0	1.0	0.0	**90.0**	0.0	0.0
8	8.1	7.1	2.0	3.0	1.0	2.0	3.0	**72.7**	1.0
9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**100.0**

4 Concluding Remarks

As we may observe, the results are considerable lower than the results presented on the Brazilian dataset. It is not necessary to mention that the Brazilian database, in spite of all variability it presents, is a laboratory database. In other words, a worst performance was expected for the Chilean database. We believe, however, that the performance can be increased when more data become available. Then we can retrain the HMM models considering all characters appearing in the Chilean month words and also make the models absorb some of the variability of the Chilean writing style. In the meanwhile, we will extract other features (concavity-based) and use them in a multiple-codebook strategy.