Influence of Word Length on Handwriting Recognition

F. Grandidier[1,2], R. Sabourin[1,2], A. El Yacoubi[1,3], M. Gilloux[4] and C.Y. Suen[1]

[1] CENPARMI, Concordia University,1455 de Maisonneuve Blvd West, Montréal, Canada, H3G 1M8
[2] LIVIA, Ecole de Technologie Supérieure, 1100 rue Notre Dame Ouest, Montréal, Canada, H3C 1K3
[3] PPGIA, Pontifícia Universidade Católica do Paraná, R. Imaculada Conceição 1155, 80215-901 Curitiba-PR, Brazil
[4] RMO, Service de Recherche Technique de La Poste, BP 86334, 44263 Nantes Cedex 02, France

Abstract

Two strategies can be considered in handwriting recognition: phrase or word approaches. In this paper we want to demonstrate the superiority of the phrase one, especially in city name recognition. The performances of an HMM-based off-line system using an analytic approach with explicit segmentation are evaluated on 2 databases: (i) city names in full and (ii) city names in single words. A difference between their performances is observed, principally caused by the dissimilarity of word lengths between the databases. After generating other data sets and lexicons, experiments were performed yielding results which lead us to conclude that word length in the data set as well as in lexicons, significantly influences recognition performance, and also that is preferable to perform city name recognition based on phrase approach than word recognition.

1. Introduction

Automatic handwriting recognition has several applications. In most of them the system must recognise a phrase or a sequence of words (cheque processing, address reading). However, the recognition is usually performed at the word level with the help of a lexicon, then a post-processing phase is carried out to validate the combination of words [1].

Among all handwriting recognition systems some classifications are usually made. The target application and the associated vocabulary size constrain system development: global [2] or analytic [3] approaches could be considered. Segmentation algorithms are also classified into 2 categories: implicit [4] and explicit [3]. The former performs *a priori* while the latter uses characteristic points to split word into graphemes. The technique to perform the recognition can be chosen from several. Recently hidden Markov models (HMM) have become one of the popular techniques used [3-7]. The recognition strategy used can also be considered: isolated words or phrases. In most current systems the first strategy is preferred.

In this paper we want to demonstrate the influence of word length on the performance of a handwriting recognition system, and thus the superiority of phrase recognition over word recognition

2. System overview

Our system is designed for the recognition of handwritten words or sequences of words such as those found on envelopes. It is an HMM-based off-line system using an analytic approach with explicit segmentation. The target application constrains it to take into account all kinds of handwriting: cursive, hand-printed and mixed.

First several preprocessing steps are performed in order to reduce noise in the input images, and to remove most of the variability of the handwriting. This stage is done in four steps: baseline slant normalisation, character skew correction, lower case letter area normalisation when dealing with cursive words, and smoothing. For more details on the preprocessing see [5].

Our system uses an analytic approach; thus a segmentation process must be carried out. As it is a very difficult task to split word into letters, our segmentation algorithm produces more segments, generally smaller than letters. During the recognition phase the model used will cluster them into letters with the help of their context.

Each segment previously obtained will be transformed into a set of two symbols each from a different set of features. The first (27 symbols) is based on global features: ascenders, descenders and loops. The second feature set (14 symbols) is based on the analysis of the horizontal and vertical contour transition histograms of each segment. In order to add contextual information to the system, the nature of segmentation points are encoded with the help of 5 features. Finally, this step allows us to represent the city name image by two feature sequences of equal length, each consisting of an alternation of symbols encoding the segment shape, followed by symbols encoding the segmentation point. The two sets of features are considered to be independent, owing to the fact that the features are independently extracted. For more details on segmentation and feature extraction phase, see [6].

The large vocabulary associated with postal application

constraints the system to use letter level modelling. In order to overcome the imperfect results of the segmentation phase (over- and under-segmentation), we use a multiple-path left-to-right HMM. It must be noted that observations are emitted along transitions, and for transitions modelling shape segments two symbols are emitted independently. A special model is also used to characterise the space between words.

During the learning phase, as we have exact labelling, the city name model is built by concatenating the appropriate elementary letter models. Then the Baum-Welch algorithm is used to estimate the best parameter values of character models. We use a training and a validation set of data during this phase, the latter in order to evaluate the improvement; this way, the system will not be specialised on the training data. For more details on the model and its use, see [7].

In the recognition phase, driven by the lexicon, city name models are built for each entry of this lexicon by concatenating letter models. Nevertheless, as no information is available on the writing style (cursive, hand-printed or mixed), both letter models (lower and upper case) are considered in parallel. During recognition, the feature sequence of the unknown city name is aligned with all lexicon entries, by the help of the Viterbi algorithm. The system is generally tested with the help of 3 lexicons (10, 100 and 1000 city names). They are randomly built by drawing city names from a global vocabulary of 6815 city names (we called it: *SRTP standard lexicon*). Now the system works without a rejection procedure, so we always include the correct city name in each lexicon. In addition to the recognition rate, we use the relative perplexity P_R to evaluate the performance of our system. This measure is related with the notion of relative entropy H_R, which is equivalent to the Kullback-Leibler distance [2], by the formula: $P_R = 2^{H_R}$. The relative entropy is given by:

$$H_R = D_{KL} = \frac{1}{m}\sum_{i=1}^{m} -\log\left(\Pr_{corr}^{i}\right)$$

where m is the number of examples in the test set, and $\Pr_{corr}^{i}$ the *a posteriori* probability of the correct class for the example i given by the recognition module. This perplexity indicator allows measuring the difficulty of the recognition task.

3. Evaluation of the system performance

The SRTP database was used to develop this system. However, we also tested it on the database provided by the "Centre of Excellence on Document Analysis and Recognition" (CEDAR). We will describe below both databases and the results obtained.

3.1. Performance on the SRTP Database

The SRTP database is composed of unconstrained handwritten French city name images manually located on real life envelopes. It is important to note that city names are considered as one entity, even if there is more than one word. The training, validating and testing sets contain respectively 12023, 3475 and 4674 city names. In order to characterise and compare databases, some statistics were calculated: the mean length in number of letters of city names in the different sets (respectively 10.7, 11.6 and 11.1), and the distribution of city names (from the 3 data sets) according to their length in characters (see Figure 1).

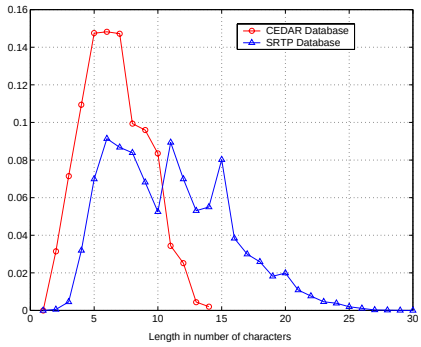

Figure 1 : Databases characterisation

The system described in section 2 was trained on the SRTP database and then tested. The results are given in Table 1. As we can deduce from the perplexity values, the recognition task is more difficult for larger lexicon size. The recognition rate confirms this remark, since it loses 12 percent between tests as lexicon size increase.

Recognition Rate			Relative Perplexity		
10	100	1000	10	100	1000
98.9	95.3	86.9	1.05	1.24	1.84

Table 1 : Results obtained on the SRTP database

3.2. Performance on the CEDAR Database

The CEDAR database [8] is composed of unconstrained handwritten data from US mail envelopes arranged in several sub-databases. Only the city name fields from the BD and BS databases are suitable for our experiments. We must note that city names consisting of more than one word are split into words in all sets (*New York City* becomes 3 examples in the different sets), and also that the Otsu algorithm was used for binarization.

In order to perform our system training, we split the pre-defined training set into two parts: training (3108 examples) and validation (529 examples). The testing set is composed of 377 words. In addition, 3 lexicons of different sizes are given for each city name of the testing set. The characterisation of the database was performed as described previously. The mean lengths of the training, validation and testing sets are respectively 6.6, 6.9 and 6.6 characters. The distribution of all words contained in the database is presented in Figure 1.

After training, testing was performed with the provided lexicon. However, as our system works without rejection,

we added the correct word to the lexicon. The results obtained are presented in Table 2. Performances are really worse than those obtained with SRTP database. With lexicon size 1000, the recognition rate falls more than 30 points. However, our results are not so far from those of others who worked on the same data set [4].

Recognition Rate			Relative Perplexity		
10	100	1000	10	100	1000
88.9	75.8	56	1.62	3.48	11.57

Table 2 : Results obtained on the CEDAR database

3.3. Reason of performance differences

First, our system was designed with the help of the SRTP database, so the letter HMM used better fits this data and preprocessing as well. For example, our system has no algorithm allowing the removal of underlining, which is frequent in the CEDAR database. Another reason is the relatively small size of this database. But the most important reason is certainly the difference between database with respect to word length (see Figure 1).

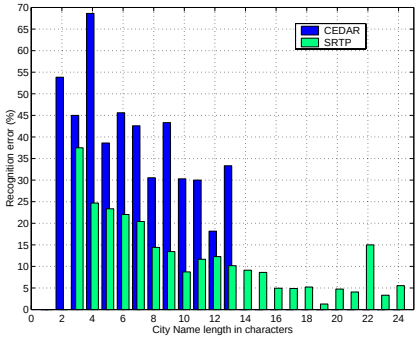

Figure 2 : Error rates according to example length

In order to demonstrate this point, some statistics were collected. The error rate was calculated for city name length from both testing sets (Figure 2). Whatever the example length, the error rate is greater for CEDAR test. For both databases, we can notice that the longer the example, the lower the error rate. This remark confirms the influence of word length on system performance. Another point is that the length in characters of the entries of the lexicon can also influence the system performance.

4. Influence of word length

Several experiments were carried out in order to verify the influence of the word length. In addition, the influence of the training set size was also studied. First, in the same way as the previous SRTP test, we constructed the curve representing the recognition rates with respect to the number of examples in the training set for the 3 lexicon sizes. These 3 curves marked with "Δ" on Figure 3 represent the standard system using city names in data sets and in lexicons.

In order to study the influence of word length on our system, the SRTP database was decomposed from city names into single words. The sizes of the training, validation and testing sets become respectively 18343, 5122 and 7034 words, with the corresponding mean lengths 5.8, 6 and 5.8 respectively. The word distributions according to length in characters are similar to the CEDAR database. However, there are more 2-character length words in the single word SRTP database, because of the French city name vocabulary (et, en, le, St...). Two new global lexicons were built in order to evaluate the influence of the lexicon entries size. The first (called *single lexicon*) contains only single words, it is the decomposition of the *SRTP standard lexicon*. The second is the concatenation of both (*standard* and *single lexicons*) called *compound lexicon*. In order to compare the performance of our system on SRTP and CEDAR databases, we also built global lexicons for CEDAR (both *single* and *compound*). Some statistics were made on these global lexicons: mean length and word distribution with respect to their length in characters. Finally we conclude that both *single lexicons* are similar, as well as both *compound lexicons*, and also that the SRTP *compound lexicon* has the same characteristics as the *SRTP standard lexicon*.

Two series of experiments were performed with the single word SRTP database. For both, the size of the training set was increased from 1500 words to the maximum, and 3 lexicon sizes were used (10, 100 and 1000). The first succession (marked "∇" on Figure 3) was performed with the help of the SRTP *single lexicon*. The second used the SRTP *compound lexicon*; the corresponding curves are marked "O". For each test 3 curves are plotted; the upper is always for lexicon size 10, and the lowest for lexicon size 1000.

Lexicons	Recognition Rate			Relative Perplexity		
	10	100	1000	10	100	1000
Provided	88.9	75.8	56	1.62	3.48	11.57
Single	88.3	70.3	49.6	1.59	4.22	18.37
Compound	90.2	77.7	57	1.43	3.20	11.93

Table 3 : Performances on CEDAR data. with different lexicons

Similar tests have been performed on the CEDAR database, to evaluate the influence of word length in the lexicon; and the results are presented in Table 3.

The influence of word length in data sets on system performances can be directly estimated by comparing the standard system curves (marked "Δ") with curves obtained from the single word SRTP database and the SRTP *compound lexicon* (marked "O"). We can notice that the recognition rate of the second series is always lower than the first, and the difference increases with lexicon size. The perplexity indicator was used to quantify this difference. For lexicon sizes 10, 100 and 1000, values show respectively an average increase of 12%, 45% and 139% by using single words in data sets instead of city names. This result leads us to conclude that there is a

significant influence of word length on the handwriting recognition task, and that it is easier to recognise long words than short ones. In the latter case, only a few segments will be generated during the segmentation process, and so only a few features will be extracted. For long words the system has more features and contextual information to perform the recognition. In addition, there is greater chance that the most discriminative features of our sets will be present in the feature sequences. So the discriminative power of each feature is very important to perform recognition of short words.

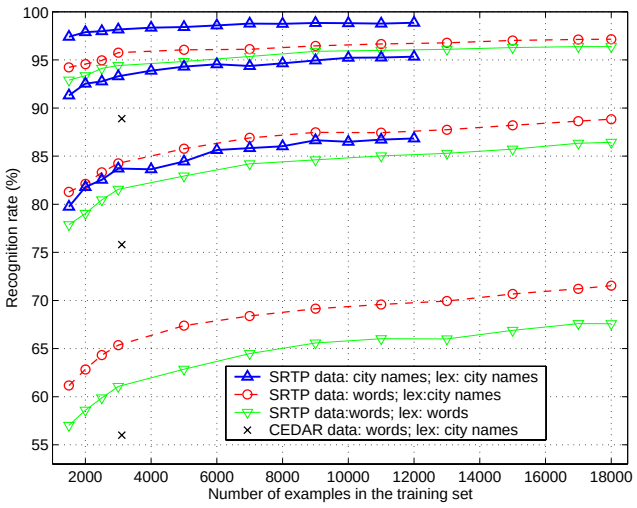

Figure 3 : All Recognition Rates

The influence of the lexicon can also be estimated directly from Figure 3 by comparing the curve marked "O" with the one marked "∇". Tests with the SRTP *single lexicon* show respectively, for sizes 10, 100 and 1000, 4%, 13% and 23% perplexity average increases with respect to the use of the SRTP *compound lexicon*. This influence can also be evaluated in the CEDAR database by comparing the 2 last lines of Table 3. The same observation as above can be made: the *compound lexicon* reduces the recognition task difficulty. The reason is that during the recognition phase the use of compound lexicons leads to bigger length dispersion among all candidates randomly drawn from the lexicon, thus there is less confusion introduced by the lexicon, and recognition becomes easier.

The system performance on SRTP and CEDAR databases, in the same experimental conditions, can be compared by referring to Figure 3 (SRTP: curves marked "O", CEDAR: points marked "×"). There is still a difference that can be attributed to the database difference (writing style, noise, underlining, data set size), and to the fact that our system was developed on the SRTP database (preprocessing and the model better fit these data).

5. Conclusion

After a description of our system, we presented its performances on 2 databases collected in 2 different continents. The results obtained lead us to evaluate the influence of word length in data sets as well as in lexicons. To achieve this goal, we decomposed the SRTP city name database into a single word database, and we generated global lexicons containing city names or single words. After several tests we concluded that word length in data sets as well in lexicons influences system performances. By using city names (phrase approach) instead of words, the performance improved. Therefore we can conclude that city name recognition is preferable to word recognition.

Finally we made some remarks on the CEDAR database. The sizes of the training and testing sets are small, so it is difficult to be confident in the results obtained from this database. A bigger one must be used to properly perform the system training and evaluation, otherwise it could memorise the database.

Acknowledgements : This work was supported by the Service de Recherche Technique de La Poste (SRTP) at Nantes, France, the Ecole de Technologie Supérieure and the Centre for Pattern Recognition and Machine Intelligence at Montréal, Canada.

References

[1] G. Kim and V. Govindaraju, "Handwritten Phrase Recognition as applied to Street Name Images," Pattern Recognition, Vol.31 (1), pp 41-51, 1998.

[2] S. Knerr, O. Baret, D. Price, and J.C. Simon, "The A2iA Recognition System for Handwritten Checks," Proc Document Analysis Systems, Malvern, Pensylvania, pp 431-494, Oct 14-16, 1996.

[3] M.Y. Chen, A. Kundu, and J. Zhou, "Off-Line Handwritten Word Recognition Using a Hidden Markov Model Type Stochastic Network," IEEE Trans. on PAMI, Vol.16 (5), pp 481-496, 1994.

[4] M. Mohamed and P. Gader, "Handwritten Word Recognition Using Segmentation-Free Hidden Markov Modeling and Segmentation-Based Dynamic Programming Techniques," IEEE Trans. on PAMI, Vol.18 (5), pp548-554, 1996.

[5] A. El-Yacoubi, R. Sabourin, M. Gilloux and C.Y. Suen, "Off-Line Handwritten Word Recognition using Hidden Markov Models," in Knowledge-Based Intelligent Techniques in Character Recognition, L.C. Jain & B. Lazzerini eds, CRC Press, pp 193-229, 1999.

[6] El-Yacoubi, A., Gilloux, M., Sabourin, R. and Suen, C.Y., "Unconstrained Handwritten Word Recognition using Hidden Markov Models," To appear in IEEE Trans. on PAMI.

[7] A. El-Yacoubi, R. Sabourin, M. Gilloux, and C.Y. Suen, "Improved Model Architecture and Training Phase in an Off-line HMM-based Word Recognition System," Proc. 13th ICPR, Brisbane, Australia, pp 1521-1525, Aug. 16-20, 1998.

[8] J.J. Hull, "A Database for Handwritten Text Recognition Research," IEEE Trans. on PAMI, Vol.16 (5), pp 550-554, 1994.