Hidden Markov model topology optimization for handwriting recognition

Núria Cirera, Alicia Fornés and Josep Lladós
Computer Vision Center
Universitat Autònoma de Barcelona, Ed. O
08193, Bellaterra, Spain
Email: {ncirera,afornes,josep}@cvc.uab.cat

***Abstract*—In this paper we present a method to optimize the topology of linear left-to-right hidden Markov models. These models are very popular for sequential signals modeling on tasks such as handwriting recognition. Many topology definition methods select the number of states for a character model based on character length. This can be a drawback when characters are shorter than the minimum allowed by the model, since they can not be properly trained nor recognized. The proposed method optimizes the number of states per model by automatically including convenient skip-state transitions and therefore it avoids the aforementioned problem. We discuss and compare our method with other character length-based methods such the Fixed, Bakis and Quantile methods. Our proposal performs well on off-line handwriting recognition task.**

I. Introduction

Handwriting recognition systems are composed by optical models to recognize characters, and language models to enhance the recognition rate. One of the most popular optical models for handwriting recognition are hidden Markov models [1], because of their early success on speech recognition and solid mathematical foundation.

Since latin-based text is written from left to right, most handwriting recognition systems use a sliding window that follows the text direction and extract a sequence of feature vectors from the input image. For this same reason models for each character are designed with left-to-right topology. Character writing variability is handled by the correspondent hidden Markov model. These model parameters are the mixture weights, transition probabilities and probability distribution parameters (usually Gaussian distributions are used).

Nevertheless, two hyper-parameters must be set before training, together with the topology or inter-state transitions placement. These are the number of states per model and the number of Gaussians per state. We focus our work on the optimization of the number of states together with the model topology.

Some studies show that topology, not number of states, has the strongest influence on recognition [2]. It is observed in [2] that left-to-right topology performs better than ergodic topology since the later has more parameters to estimate. Most handwriting recognition approaches impose the topology design independently on the task or data at hand. On the other side, there are approaches based on the observed features values, minimizing objective functions such as model likelihood or entropy. On the other side, other approaches use character geometric constrains or statistics to infer the number of states for each model.

Since likelihood is a non-decreasing function of the number of states, data-driven methods have to leverage likelihood with a penalty term in order to keep a simple yet competitive model. Bayesian Information Criterion (BIC) penalizes long models. In [3] BIC method was used to select the number of states, fixing the number of Gaussians to five and using a fully connected topology for all models. Compared with other methods, accuracy gain is minimal but simpler models are achieved. Another type of penalty was proposed in [4] where discriminative power among models is maximized together with the likelihood. Their Discriminative Information Criterion achieves a slightly higher recognition rate than BIC, at the expense of more complicated models.

Some on-line handwriting recognition systems apply methods related to character geometric constrains. For instance, in [5] number of states is proportional to the number of strokes per character. In the case of off-line handwriting recognition this stroke information is not available, but the approaches are similar as the ones we describe in the following paragraphs.

Other approaches investigate the iterative addition of states, others the deletion or merging of states. In [6] each model is iteratively compared with the ones with one state more and one state less, and afterwards the best model with respect to likelihood and recognition rate is selected. Another interesting and more recent proposal in [7] iteratively deletes model states based on the inter-state transition probabilities. For this, different behaviours are detected on each state and next iteration reduces the number of states accordingly.

Most methods for number of states selection in off-line handwriting recognition tasks use the character length histogram. Fixing all models to have the same number of states is very popular, although it usually results in lower recognition rates [8][9] when compared to more sophisticated state-of-the-art methods. Another method proposed in [10], called Bakis method, assigns the mean observed character length as the number of states. It outperforms the Fixed method by far while keeping relatively short models [8]. A linear combination of both Fixed and Bakis methods is proposed in[11] for on-line recognition. It is based on imposing a lower borderline in the number of state selection made by Bakis method. Finally, Quantile method [8] is based on the quantile of the character length histogram. For each model it assigns the number of states to be the character length below which a certain amount of observed lengths fall. It performs similarly with Bakis but brings much more interpretability to the results. However, the recognition decay of Quantile method shows that models longer than the majority of observed character lengths perform poorly.

When modeling characters with high length variability, if the methods chooses a low number of states then many long instances will not be properly trained nor recognized. Contrary, if the number of states is too high, such model is not able to properly recognize short instances of that character. Thus this is a critical decision for the above-mentioned methods, since they do not provide a criteria for making flexible this hard decision when defining the rest of characteristics of the model.

Therefore, our aim is to benefit from both long and short alternatives in a flexible way. We believe that the addition of convenient skip-state transitions based on the character length histogram can improve the recognition rate of long models.

In this paper we propose a method to optimize the number of states together with the topology, based on Quantile method. We compare it with three state-of-the-art methods, these being Fixed, Bakis and Quantile methods. We describe them in Section II. This is followed by a brief discussion on number of state selection and our proposed method in Section III. Finally, all methods are evaluated and compared in Section IV and some conclusions and future work are drawn in Section V.

II. State-of-the-art Methods

A. Fixed Length Modeling

Let C be the finite set of characters and NS_c be the number of states of a hidden Markov model for character c. A straightforward method to determine NS_c is setting all models to have the same number of states (see Equation 1).

$$NS_c(\alpha) := \alpha \quad \forall c \in C, \;\; \alpha \in \mathbb{N} \tag{1}$$

This method might overestimate the number of states for short characters such as i and underestimate the number of states for long characters such as m. However it is one of the most popular methods due to its simplicity.

B. Bakis Length Modeling

In this method [10] each character is assigned an integer fraction of its average sampled length (see Equation 2). In this way, the real-valued α parameter allows to assign a different number of states to each character model, so longer characters such as m may have longer models, and vice-versa for short characters.

$$NS_c(\alpha) := \left\lceil \alpha \cdot \sum_{i=1}^{n} x_i/n \right\rceil, \;\; \alpha \in [0,1] \tag{2}$$

A major drawback comes from its sensitivity to extremal values of the observed character lengths. Since NS_c is a linear combination of the observed lengths mean, the outlier observations can drag the mean away from the most observed length, this being the mode. Apart of its sensitivity to outliers, the mean length value does not convey information regarding the whole set of observed lengths. This makes difficult to interpret the effect of α parameter on the selection of NS_c as well as its relation with the entire set of lengths.

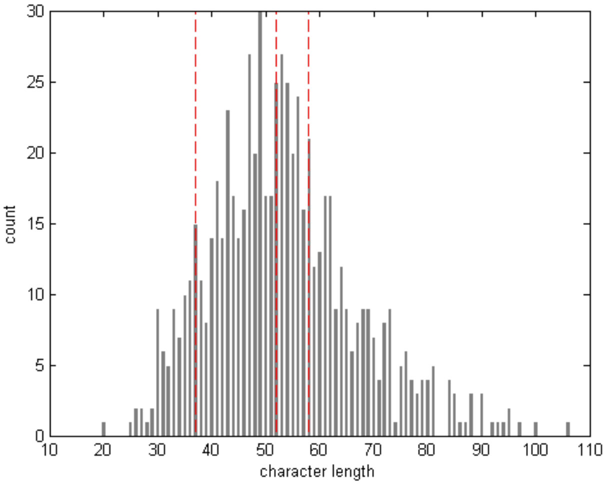

Fig. 1. Example of α parameter effect in Quantile method for character m length histogram. Vertical dashed lines mark the number of states that quantile method selects for α = 0.1, 0.5 and 0.7 (from left to right respectively).

Therefore, although this method overcomes the drawbacks of Fixed method, it is sensitive to outliers and there is no guarantee NS_c represents most sampled lengths.

C. Quantile Length Modeling

Quantile length modeling [8] assigns NS_c to be the $100 \cdot \alpha$-th percentile of the observed lengths for character c (see Equation 3). In other words, it retrieves the sampled length value below which the $100 \cdot \alpha$ per cent of samples fall (see Figure 1).

$$NS_c(\alpha) := \max_{x_i}(P(X < x_i) \leq \alpha), \;\; \alpha \in (0,1) \tag{3}$$

In comparison with Fixed and Bakis methods, it selects NS_c by using a parameter that has a comprehensive role in it, as one can see depicted in Figure 1. For instance, setting α to 0.7 means that the resulting NS_c does not reflect he 70% of the shorter observed lengths.

III. Proposed Quantile-based hist2NSkip Method

A. Length Modeling Discussion

In the previous section three state-of-the-art methods for model length selection have been reviewed. They all rely on a single α parameter. For these methods, small values of α generate short models, and this is desirable in many scenarios.

In segmented character recognition, shorter images will be misclassified by long models, since its feature vectors can not be modeled by such hidden Markov model topology. For the same reason short images can not be used for training as well. This means shorter models are required, since they can be trained with more images, and therefore better estimates for the model parameters can be obtained.

In another tasks such as segmented word recognition, a longer model for character c, applied on a word image where c is shorter, may use its starting and ending states to incorporate surrounding features from other characters in these dummy states. So yet again, shorter models are also preferred for this task.

The number of states in a hidden Markov model is related to the horizontal complexity of the target character. In practice it has been observed that low number of states are not optimal as well [8]. There is an increasing performance that after a certain point decreases on α. Hence too short models or too long models do not perform optimally. In the former case, a simplistic hidden Markov model can not reflect the variability from the training images. In the later case, a longer hidden Markov model can either become unable to generalize to new text images due to overfitting, or suffer from poorly estimated parameters since too many images lay below the proper training range [2].

We propose to improve performance of longer models by avoiding discarding a large amount of shorter images.

B. Histogram-based Skip-state Definition Method

Out of the three state-of-the-art methods discussed in Section II, Quantile method is the only one that provides information on the histogram of character length in a controlled manner. The proposed method takes advantage of the Quantile method and adds convenient skip-state transitions as we describe in this subsection.

The method optimizes the number of states NS_c and determines the number of skip-states $NSkip_c$ for each model c. This method is summarized in Algorithm 1 and described in detail in the following paragraphs.

Algorithm 1 hist2NSkip Algorithm

```
function HIST2NSKIP($\{x_i\}, \alpha$)    ▷ $\{x_i\}$ observed lengths
    $NSkip_c \leftarrow 0$    ▷ Initialize variable
    $NS_c \leftarrow Quantile(\{x_i\}, \alpha)$    ▷ Quantile method
    if even($NS_c$) then
        $maxNSkip_c := NS_c/2 - 1$
    else    ▷ $NS_c$ odd
        $maxNSkip_c := \lfloor NS_c/2 \rfloor$
    end if
    $minNE \leftarrow NS_c - maxNSkip_c$
    sortedX $\leftarrow$ sort($\{x_i\}$)    ▷ Ascending sorting
    $aux \leftarrow$ sortedX[1]    ▷ Shortest length
    $NSkip \leftarrow maxNSkip_c$
    while $aux < minNE$ do
        sortedX $\leftarrow$ sortedX[2 : $end$]    ▷ Drop 1st
        $aux \leftarrow$ sortedX[1]
    end while
    $NSkip_c \leftarrow NS - aux$
    return $NSkip_c$
end function
```

First, Quantile method is applied. Based on the character length histogram it provides the number of states NS_c for the linear left-to-right model for character c. Once the number of states is set, based on this topology, there is an upper bound on the number of skip-states transitions named $maxNSkip_c$ that can be placed in the model design. Consequently, for each $NSkip = 0, 1, \ldots, maxNSkip_c$ there is a minimum amount of observations produced when traversing the model, named $minNE_c$ (see Figure 2).

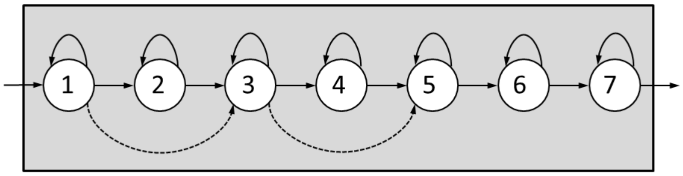

Fig. 2. Example of hidden Markov model with 7 emitting states connected from left to right with self-loops, and two skip-state transitions connecting state 1 to state 3, and state 3 to state 5. It allows a minimum number of 5 emissions. Following the notation, NS_c=7 and $NSkip$=2 sets $minNE_c$=5.

Fig. 3. Three instances of character m in NIST database.

We aim to build a model that can be traversed in less than NS_c transitions if data requires so. Hence the topology we design would have the first hypothetical skip-state at the top-left state, given that state transitions are left-to-right with self-loops. Then every two states a hypothetical skip-state would be considered (see Figure 2). Following this self-imposed criteria for skip-state placement, $minNE_c := NS_c - NSkip$.

We then select the minimum observed length of c character that falls into the range defined by $minNE$ when $NSkip = 0, 1, \ldots, maxNSkip_c$. For this we just need to sort the length histogram in ascending order and iteratively check if the shortest length in the iteration can be produced by a model with NS_c states with 1 to $maxNSkip_c$ skip-state transitions, until a certain length actually fulfils this condition. This procedure univocally defines $NSkip_c$ and actual skip-states are finally placed accordingly in the hidden Markov model.

IV. Experimental Evaluation

Corpus We have evaluated the performance of Fixed, Bakis, Quantile and hist2NSkip methods on a subset of NIST database [12]. This database contains images of handwritten characters and numbers from different writers (see Figure 3). The subset we have used covers the 26 low case English characters from a to z [1]. A total of 17,898 images are used for training and 8,951 for testing, randomly selected for this partition.

Model Topologies All models used in Fixed, Bakis and Quantile experiments have a linear left-to-right topology with no skip-state transition. All models of hist2NSkip method have some skip-state transitions placed according to III-B. All state

[1] The methods we compare rely on character length. Hence we preferred to avoid evaluation on word based datasets in order to avoid further outlier values due to poor character alignments.

emission probabilities are modeled with one Gaussian [2].

Feature Extraction Original images are 128x128 pixels and each one contains a single character. Images are segmented based on vertical projections in order to discard the initial and final blank areas. For each image a sliding window of 1 pixel width in 1 pixel steps is applied from left to right. A feature vector is extracted from each window. We use the well-known nine geometrical features from [13], which are the 0, 1st and 2nd moments, upper and lower foreground pixels' distance to the upper and lower borders respectively, gradient of the contour at these positions, number of background to foreground transitions, and finally the average pixel density between the top and bottom foreground pixels.

Training and Decoding Algorithms Model parameters are estimated using four iterations of the Baum-Welch algorithm. The models for each character are initialized based on the character class global mean and variance. Recognition is done by the Token-passing algorithm.

Software We used Matlab R2012a to preprocess the images, extract features, implemented Fixed, Bakis, Quantile and hist2NSkip methods and generate the figures included in this paper. We use the HTK Toolkit for model training and recognition.

Results Each one of the four methods described in sections II and III-B are evaluated on the test set. Recognition rate is reported in Tables I to IV, together with the amount of model parameters. The best recognition rate for each method are in bold.

Comparing the performance of Fixed method versus the others, we can see that Bakis, Quantile and hist2NSkip are much superior. Methods that adapt the number of states to each character achieve a good trade-off in terms of model efficiency vs. performance.

Bakis method improves this trade-off with respect to Fixed method, but does not achieve a competitive performance when compared with both Quantile-based methods.

It is very interesting to observe that for small values of α (i.e. for shorter models), Quantile method outperforms hist2NSkip method. For instance, when α=0.02, the recognition rate of Quantile method is 69.86%, while the proposed method only achieves a 67.80%. Therefore state-skip transitions are not convenient for short models.

However this trend is reversed for α larger than 0.10, when our proposal outperforms the original Quantile method. For instance, when α=0.20 Quantile method achieves a 56.68% whereas our method achieves 70.95%. This means that for medium to long models adding convenient state-skip transitions is beneficial, since the recognition increases while keeping similar number of parameters. Indeed, the increase in the amount of parameters for α=0.20 is marginal (13,680 versus 13,982), whereas the performance improvement is noticeable. Furthermore, the best result (70.95%) is obtained using the proposed Quantile-based hist2NSkip method, which is indeed a long model.

TABLE I. RESULTS OF FIXED METHOD.

α	Rec. rate (%)	# Par.
2	43.06	1,040
4	55.55	2,080
6	55.37	3,120
8	57.65	4,160
10	59.30	5,200
12	59.77	6,240
14	61.26	7,280
16	**61.66**	8,320
18	60.59	9,360
20	61.17	10,400
22	60.32	11,440
24	57.14	12,480

TABLE II. RESULTS OF BAKIS METHOD.

α	Rec. rate (%)	# Par.
0.2	59.87	3,780
0.30	63.14	5,600
0.36	65.52	6,740
0.38	66.65	7,060
0.40	66.31	7,420
0.42	67.11	7,800
0.44	67.24	8,120
0.50	**67.47**	9,160
0.60	66.15	10,980
0.70	59.94	12,760

TABLE III. RESULTS OF QUANTILE METHOD.

α	Rec. rate (%)	# Par.
0	66.63	6,760
0.01	69.36	8,860
0.02	**69.86**	9,620
0.03	69.14	10,100
0.04	68.71	10,440
0.05	67.72	10,780
0.10	64.15	12,020
0.20	56.68	13,680
0.50	34.15	17,220
0.60	27.54	18,440
0.70	19.92	19,920

TABLE IV. RESULTS OF HIST2NSKIP METHOD.

α	Rec. rate (%)	# Par.
0	66.63	6,760
0.01	68.83	8,957
0.02	67.80	9,755
0.03	67.46	10,257
0.04	67.48	10,614
0.05	67.43	10,971
0.10	68.80	12,260
0.20	**70.95**	13,982
0.50	70.00	17,629
0.60	68.04	18,879
0.70	63.17	20,399

V. CONCLUSION

In this paper a method to select the hidden Markov model topology is proposed. First the numbers of states are fixed based on Quantile method. Then a convenient amount of skip-state transitions are inferred from the character length histogram and included in the model. In this way the model can consider shorter feature sequences.

Based on the evaluation of this proposed method hist2NSkip, we see that adding skip-state as described in Algorithm 1 improves the recognition rate over the Fixed, Bakis and Quantile methods in most cases. It is specially convenient when defining medium and long models. Since our training and evaluation is on segmented characters, the performance improvement of our method over the original Quantile method is the result of increasing the training set.

[2] In preliminar experiments we observed that the performance when varying the number of states has the same trend for whatever number of Gaussians is set. Therefore, we decided to model each state with one Gaussian in order to ease comparisons with [8]

We are able to train over more images since the skip-state transitions allow for shorter feature vectors to be used in the training.

In the future we plan to evaluate our proposed method for word recognition, where we expect similar results. We may apply some preprocessing to correct the skew and the slant in order to start with a more controlled data.

Furthermore, we also want to explore the performance of our method when increasing the number of Gaussians per state. In this case we foresee similar relative performance among methods. In this paper we proposed a method to optimize the number of states and number of skip-sates transitions altogether. We may try to also optimize the number of Gaussians in the same method as in [3] and [9].

ACKNOWLEDGMENT

This work has been partially supported by the Spanish project TIN2012-37475-C02-02 and the European project ERC-2010-AdG-20100407-269796.

REFERENCES

[1] T. Plotz and G. A. Fink, “Markov models for offline handwriting recognition: A survey,” *Int. J. Doc. Anal. Recognit.*, vol. 12, no. 4, pp. 269–298, Nov. 2009. [Online]. Available: http://dx.doi.org/10.1007/s10032-009-0098-4

[2] K. T. Abou-Moustafa, M. Cheriet, and C. Y. Suen, “On the structure of hidden markov models,” *Pattern Recognition Letters*, vol. 25, no. 8, pp. 923–931, 2004. [Online]. Available: http://dx.doi.org/10.1016/j.patrec.2004.02.005

[3] D. Li, A. Biem, and J. Subrahmonia, “Hmm topology optimization for handwriting recognition,” in *Acoustics, Speech, and Signal Processing, 2001. Proceedings. (ICASSP ’01). 2001 IEEE International Conference on*, vol. 3, 2001, pp. 1521–1524 vol.3.

[4] A. Biem, “A model selection criterion for classification: Application to HMM topology optimization,” in *7th International Conference on Document Analysis and Recognition (ICDAR 2003), 2-Volume Set, 3-6 August 2003, Edinburgh, Scotland, UK*, 2003, pp. 104–108. [Online]. Available: http://dx.doi.org/10.1109/ICDAR.2003.1227641

[5] J. J. Lee, J. Kim, and J. H. Kim, “Data driven design of hmm topology for on-line handwriting recognition,” in *In the7th International Wrokshop on Frontiers in Handwriting Recognition*. World Scientific Publishing Company, 2000, pp. 107–121.

[6] M.-P. Schambach, “Model length adaptation of an hmm based cursive word recognition system,” in *Document Analysis and Recognition, 2003. Proceedings. Seventh International Conference on*, Aug 2003, pp. 109–113 vol.1.

[7] Z. Jiang, X. Ding, L. Peng, and C. Liu, “Analyzing the information entropy of states to optimize the number of states in an hmm-based off-line handwritten arabic word recognizer,” in *Pattern Recognition (ICPR), 2012 21st International Conference on*, Nov 2012, pp. 697–700.

[8] M. Zimmermann and H. Bunke, “Hidden markov model length optimization for handwriting recognition systems,” in *Frontiers in Handwriting Recognition, 2002. Proceedings. Eighth International Workshop on*, 2002, pp. 369–374.

[9] S. Gunter and H. Bunke, “Optimizing the number of states, training iterations and gaussians in an hmm-based handwritten word recognizer,” in *Document Analysis and Recognition, 2003. Proceedings. Seventh International Conference on*, Aug 2003, pp. 472–476 vol.1.

[10] R. Bakis, “Continuous speech recognition via centisecond acoustic states,” *Acoustical Society of America Journal*, vol. 59, p. 97, Jan. 1976.

[11] J. Geiger, J. Schenk, F. Wallhoff, and G. Rigoll, “Optimizing the number of states for hmm-based on-line handwritten whiteboard recognition,” in *in 2010 12th International Conference on Frontiers in Handwriting Recognition. IEEE*, 2010, pp. 107–112.

[12] P. J. Grother, “Nist special database 19 handprinted forms and characters database,” *National Institute of Standards and Technology*, 1995.

[13] U.-V. Marti and H. Bunke, “Using a statistical language model to improve the performance of an hmm-based cursive handwriting recognition system,” *IJPRAI*, vol. 15, no. 1, pp. 65–90, 2001.