Hidden Markov Models for Spatio-Temporal Pattern Recognition and Image Segmentation

Brian C. Lovell
The Intelligent Real-Time Imaging and Sensing (IRIS) Group
The School of Information Technology and Electrical Engineering
The University of Queensland, Australia QLD 4072
lovell@itee.uq.edu.au

Abstract

Time and again hidden Markov models have been demonstrated to be highly effective in one-dimensional pattern recognition and classification problems such as speech recognition. A great deal of attention is now focussed on 2-D and possibly 3-D applications arising from problems encountered in computer vision in domains such as gesture, face, and handwriting recognition. Despite their widespread usage and numerous successful applications, there are few analytical results which can explain their remarkably good performance and guide researchers in selecting topologies and parameters to improve classification performance.

Keywords: Image segmentation, pattern recognition

1 Introduction

There is an enormous volume of literature on the application of hidden Markov Models (HMMs) to a broad range of pattern recognition tasks. The suitability and efficacy of HMMs is undeniable and so they are established as one of the major workhorses of the pattern recognition community. Yet, when one looks for papers which address fundamental problems such as efficient learning strategies or analytically determining the most suitable architectures for a given problem, the number of significant papers is greatly diminished. So despite the enormous usage of HMMs since their introduction in the 1960's, we believe that there is still a great deal of unexplored territory.

Much of the application of HMMs is firmly based on the methodology popularised by Rabiner *et al* (1983) [14] [11] [15] for speech recognition and these studies are still a primary reference for HMM researchers. Here the Baum-Welch [3] algorithm (a version of the famous Expectation-Maximisation algorithm) is the primary tool for learning HMMs from observation sequences. However, in the words of Stolke and Omohundro [17], the Baum-Welch algorithm is far from foolproof since it uses what amounts to a hill-climbing procedure that is only guaranteed to find a local likelihood maximum. Results are very dependent on the initial values chosen for the HMM parameters.

In this paper, we evaluate several other approaches to HMM parameter estimation that yield superior results upon smaller training sets. Yet we show that these advantages do not appear to always translate to better performance on real-world pattern recognition data. We then describe a technique for using HMM related techniques for image segmentation. Along the way, we also describe the video gesture recognition system that we are using as a testbed to perform real-world evaluation of HMM learning algorithms.

1.1 Background and Notation

A hidden Markov model ([14], Chapter 6) consists of a set of N nodes, each of which is associated with a set of M possible observations. The parameters of the model include an initial state

$$\pi = [p_1, p_2, p_3, ..., p_N]^T$$

with elements $p_n,\ n \in [1, N]$ which describes the distribution over the initial node set, a transition matrix

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1N} \\ a_{21} & a_{22} & \dots & a_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ a_{N1} & a_{N2} & \dots & a_{NN} \end{pmatrix}$$

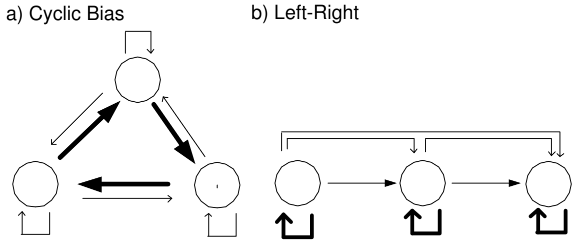

Figure 1. Cyclic and Left-Right structures. Bold arrows indicate high probability transistions. No arrow between vertices indicates a forbidden (zero-probability) transition.

with elements a_{ij} with $i, j \in [1, N]$ for the transition probability from node i to node j conditional on node i, and an observation matrix

$$\mathbf{B} = \begin{pmatrix} b_{11} & b_{12} & \dots & b_{1M} \\ b_{21} & b_{22} & \dots & b_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ b_{N1} & b_{N2} & \dots & b_{NM} \end{pmatrix}$$

with elements b_{im} for the probability of observing symbol $m \in [1, M]$ given that the system is in state $i \in [1, N]$. Rabiner and Juang denote the HMM model parameter set by $\lambda = (\mathbf{A}, \mathbf{B}, \pi)$.

The model order pair (N, M) together with additional restrictions on allowed transitions and emissions defines the topology or structure of the model (see figure 1 for an illustration of two different transition structures).

2 Comparison of Methods for Robust HMM Parameter Estimation

Four HMM parameter estimation methods were evaluated and compared by using a train and test classification methodology. For these binary classification tests we created two random HMMs and then used these to generate the test and training data sequences. For normalisation, we ensured that each sequence could be correctly recognised by its true model; thus the true models obtain 100% classification accuracy on the test data by construction. This random model generation and evaluation process was repeated 16 times for each data sample to provide meaningful statistical results.

We compared traditional Baum-Welch with ensemble averaging introduced by Davis and Lovell [8] based on ideas presented by Mackay [12], Entropic MAP introduced by Brand [4], and Viterbi Path Counting which is a special case of Stolke and Omhundro's Best-First algorithm [17]. The results in figure 2 show that these alternate HMM learning methods classify significantly better than the well-known Baum-Welch algorithm and require less training data. The Entropic MAP estimator performs well but surprisingly the performance is much the same as simple ensemble averaging which involved training multiple models using the Baum Welch algorithm and then simply averaging the models without regard to structure. Note that for a single sequence, ensemble averaging is identical to Rabiner and Juang's [14] Baum-Welch algorithm applied to multiple sequences. The best performer overall was the VPC algorithm. This method converges to good models rapidly and has been superior to the other methods in all our simulated HMM data comparisons.

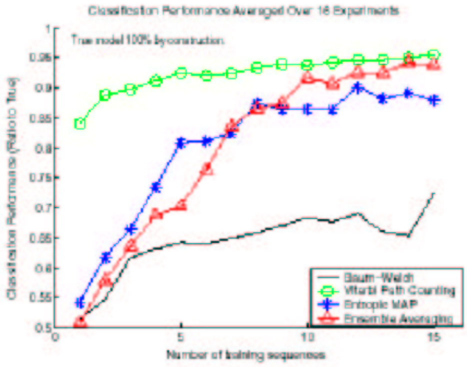

Figure 2. Relative performance of the HMM parameter estimation methods as a function of the number of training sequences. Viterbi Path Counting produces the best quality models with a much smaller number of training iterations.

3 Video Gesture Recognition

To validate the above results on HMM learning techniques on a real-world application, we developed a system for real-time video gesture recognition based on letters of the alphabet traced in space in front of a video camera. The motivation for this study is to produce a way of typing messages into a camera-equipped mobile phone or PDA using video gestures instead of attaching a cumbersome keyboard or using a pen interface. The gestures are based on single stroke letter gestures already widely used for pen data entry in PDAs. For example, Figure 3 show the hand gestures for the letters "Z" and "W."

We evaluated recognition performance over all 26 character gestures using both fully connected (FC) and left-right (LR) model topologies with the number of states

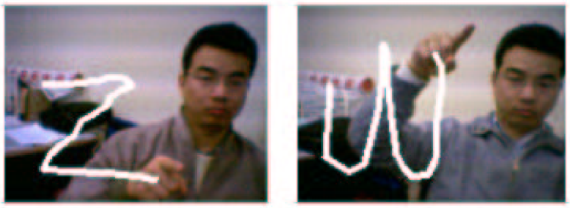

Figure 3. “Fingerwriting:” Single stroke video gesture for letters “Z” and “W.”

ranging from 4 to 10. Our video gesture database contained 780 video gestures with 30 examples of each gesture. Recognition accuracy was measured using threefold cross-validation where 20 gestures were used for training and 10 for testing in each partition.

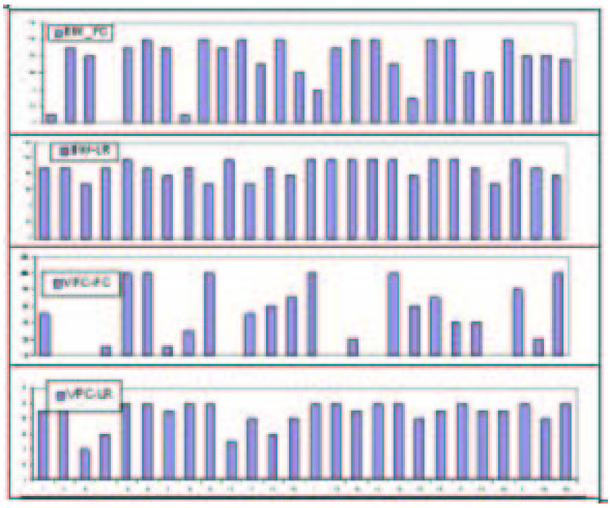

Figure 4. “Recognition accuracy for each letter gesture with number of states equal to 9. Both FC and LR topologies were tested using Baum-Welch and VPC training algorithms.

The best average recognition accuracy achieved was 90% when Baum-Welch was used for training, topology was LR, and the number of states was 9. The VPC algorithm only achieved 86% under the same conditions. Note that there is some confusion between the characters “O”, “C”, and “G” because of the similarity of the gestures. Recognition performance could be improved by 1) altering the gestures to be more distinctive, or 2) using digram or trigram letter context to improve recognition based on previous letters already recognised [16].

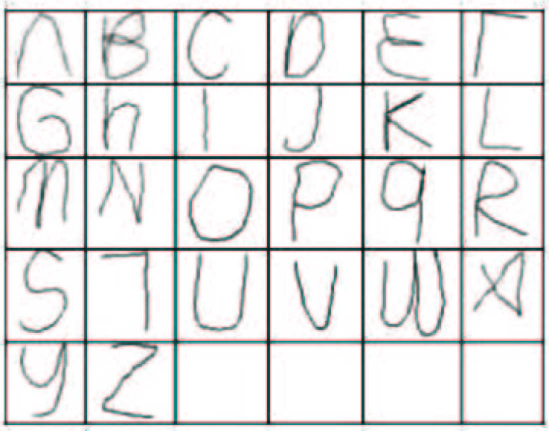

Figure 5. The alphabet of single-stroke letter hand gestures.

3.1 Discussion on Disparity in Results

Based on our simulated data studies, we had expected methods such as VPC to perform significantly better than Baum-Welch, it turned out that similar levels of recognition accuracy were obtained for all parameter estimation methods. The question that arises is, “Why does the Baum-Welch algorithm perform so well on real-world data despite its theoretical flaws and poor performance on simulated HMM data?”

A possible explanation is that this particular spatio-temporal recognition task is relatively easy, so all methods can do quite well. Unfortunately, unlike simulated data, the effort of gathering very large and diverse databases of real-world pattern recognition problems to evaluate the performance of different training algorithms is immense. Thus it is difficult to use multiple real data sets in these studies.

Unlike the random HMMs used in the simulations, McCane and Caelli [13] suggest that there are many real-world applications of HMMs that do not use the full descriptive power of the HMM. In many cases, the observation matrix B seems to provide most of the recognition performance and recognition is only weakly affected by the transition matrix A. Indeed, each row of the B matrix may be interpreted as the probability mass function of the observation symbols for a given state. Thus for a single state HMM, the B matrix would degenerate to a single row and application of the forward algorithm for recognition would be equivalent to a MAP classifier. Another interesting case is a banded LR HMM where only self-transitions and next state transitions are allowed. Thus the A matrix is of the form:

$$\mathbf{A} = \begin{pmatrix} a_{11} & 1-a_{11} & 0 & \dots & 0 \\ 0 & a_{22} & 1-a_{22} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & a_{NN} \end{pmatrix} \qquad (1)$$

In this case the expected number of observations, $\bar{n}$, (*i.e.,* duration) in state i is simply given by [15]

$$\bar{n} = \frac{1}{1-a_{ii}}.$$

Thus we can interpret A as being an adjustable clock that ideally synchronises the changes in observation statistics with changes in state to produce a time-variant MAP classifier. In many ways this HMM topology can be considered a form of dynamic time-warping — a technique used in speech recognition for many years.

Based on this intuition, we then repeated the evaluation using the banded LR model. We are pleased to report that we obtained much higher recognition accuracies of 97.3% and 96.5% with banded LR models using VPC and Baum-Welch training algorithms respectively. For a given number of states, VPC generally yielded superior models to Baum-Welch. This now confirms the trend from synthetic data studies and indicates that simple HMM topologies often work the best on real data.

4 Image Segmentation via HMMs

In this section we describe earlier work on a highly successful cell image segmentation algorithm based on active contour methods. Although we were not aware of the fact at the time, we now realise that this method can be reformulated as an example of using an HMM for image segmentation.

4.1 Cell Image Segmentation via Shortest Path Methods

The use of active contours is well established, but it is well known that these methods tend to suffer from local minima, initialisation, and stopping criteria problems. Fortunately global minimum energy, or equivalently shortest-path, searching methods have been found which are particularly effective in avoiding local minima problems due to the presence of the many artefacts often associated with medical images [6][7][9]. Here, an energy minimization method was implemented based upon a suggestion in [10]. A circular search space is first defined within the image, bounded by two concentric circles centralised upon the approximate centre of the nucleus found by an initial rough segmentation technique (*e.g.,* converging squares algorithm). This search space is sampled to form a circular trellis by discretising both the circles and a grid of evenly-spaced radial lines joining them (figure 6).

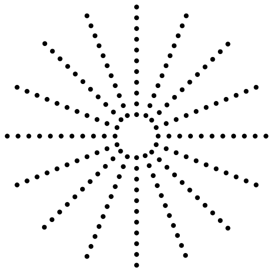

Figure 6. Discrete search space

Every possible contour that lies upon the nodes of the search space is then evaluated and an associated energy or cost function is calculated. This cost is a function of both the contour's smoothness and how closely it follows image edges. The relative weighting of the cost components is controlled by a single regularisation parameter, $\lambda \in [0, 1]$. By choosing a high value of λ, the smoothness term dominates, which may lead to contours that tend to ignore important image edges. On the other hand, low values of λ allow contours to develop sharp corners as they attempt to follow all high gradient edges, even those which may not necessarily be on the desired objects edge. Once every contour has been evaluated, the single contour with least cost is chosen as the global solution. The well-known Viterbi algorithm provides an efficient method to find this global solution as described in [2].

A data set of 19946 Pap stained cervical cell images was available for testing. These images were of the order of 128x128 pixels, quantised to 256 gray levels and each contained a single nucleus.

The single parameter λ was empirically chosen to be 0.7 after trial runs on a small sub-set of the images. This sub-set was made up of 141 known 'difficult' images from previous studies [2][1], augmented by a random sample of 269 images from the remaining data set. This careful data selection was necessary as previous experience showed that for the majority of images, the resulting segmentation was usually insensitive to the choice of λ, making the choice of optimum value difficult. Nevertheless, more demanding images require some adjustment to the parameter to achieve correct segmentation. The effect of the choice of λ on segmentation accuracy on this trial set is shown by the graph of figure 7.

With λ set at 0.0, the smoothness constraint is completely ignored and the point of greatest gradient is chosen along each search space radius. Previous studies [1] have shown that for approximately 65% of images, all points of greatest gradient actually lie upon the nucleus cytoplasm

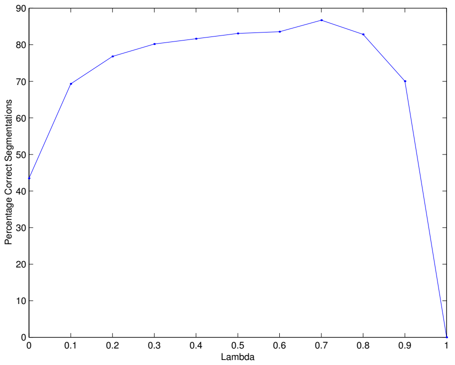

Figure 7. Plot of percentage of correct segmentations against λ for a set of images consisting of known 'difficult' images and randomly selected images.

border (figure 8(a)), so these cell images will be correctly segmented.

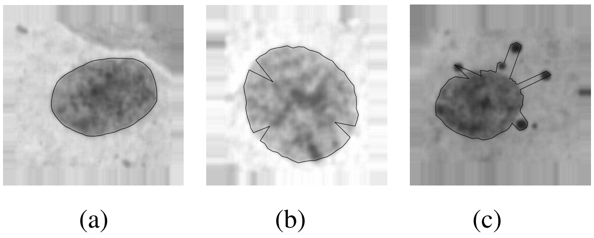

Figure 8. $\lambda = 0.0$. a) Largest gradients occur on the nucleus border, b) darkly stained chromatin generates largest gradients, c) dark artefacts generate largest gradients.

For the remaining 35% of images, a large gradient due to an artefact or darkly stained chromatin will draw the contour away from the desired border (figures 8(b)&(c)). As λ increases, the large curvatures present in these configurations become less probable (figure 9).

The graph shows a value of $\lambda = 0.7$ as the most suitable for these particular images. Every image in the data set was then segmented at $\lambda = 0.7$ and the results verified by eye. Of the 19946 images, 99.47% were found to be correctly segmented.

4.2 Reformulation as an HMM Problem

The above method can be considered as equivalent to the problem of estimating the most likely state sequence of an simple HMM from a single observation with the Viterbi algorithm. In the reformulation, the states represent the true radii of the cell and correspond to the nodes of the linearized trellis. The notion of smoothness can be embodied by forming an A-matrix such that self-transitions have the highest probability with transitions to nearby states being penalised according to smooth function of distance.

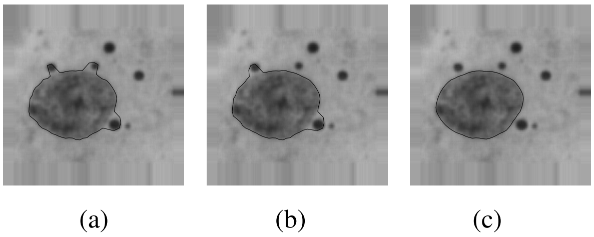

Figure 9. The effect of increasing λ. a) $\lambda = 0.1$, b) $\lambda = 0.2$, c) $\lambda = 0.5$.

The observations are the cross-sections of the gradient image as we progress around the cell image. One approach to handling these observations would be to vector quantise these observation vectors to form a discrete alphabet, so we can estimate the observation matrix, B. However, we note that the observation matrix only enters the HMM parameter estimation problem when we need to evaluate the probability of the state S_i conditioned on observation m denoted $P(S_i|O_m)$. This suggests that we can treat the cross-section of the gradient image as a likelihood function and thus obtain $P(S_i|O_m)$ directly.

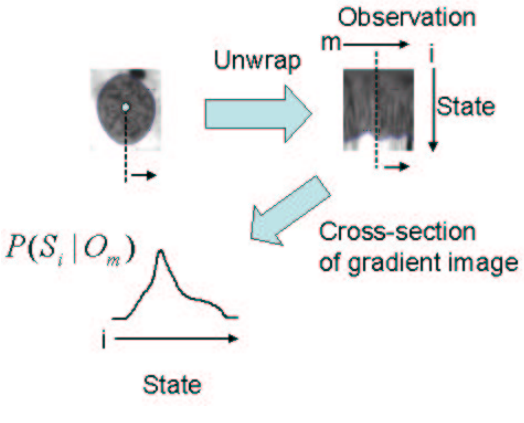

Figure 10. Gradient cross-section as a likelihood function.

The reformulation suggests that advantages may arise by being able to trade off the recognition power of the A matrix (a-priori knowledge) versus the B matrix (observed data) as we have done in this study by trading off the smoothness constraint against image gradient information. Such a tradeoff is not available in current HMM implementations, but this example gives an indication of how this may be accomplished. Such a feature could be immensely useful in the case of object shape recognition when part of the object is obscured — a extremely common occurrence in Computer Vision.

5 Conclusions

HMMs are an immensely powerful tool for solving pattern recognition and classification problems. Many studies demonstrate that it is a powerful technique, but few studies give any insight into why the performance is so good. Our own HMM training algorithm comparisons based on simulated HMM classification data do not necessarily translate into real-world data performance. We believe that that is possibly because many important real-world problems have little need for highly complex HMMs. We have also shown that HMMs are related to active contours as used for image segmentation. We hope that by unifying several pattern recognition and computer vision techniques we may gain useful insight into the design more effective algorithms.

6 Acknowledgements

This paper covers a number of research themes currently being studied within the Intelligent Real-Time Imaging and Sensing Group. I would like to acknowledge the contributions of Terry Caelli, Peter Kootsookos and my current and former students Richard Davis, Nianjun Liu, Pascal Bamford, and Christian Walder.

References

[1] P. Bamford and B. Lovell. Improving the robustness of cell nucleus segmentation. In P. H. Lewis and M. S. Nixon, editors, *Proceedings of the Ninth British Machine Vision Conference, BMVC '98*, pages 518–524. University of Southampton, September 1998.

[2] P. Bamford and B. Lovell. Unsupervised cell nucleus segmentation with active contours. *Signal Processing Special Issue: Deformable Models and Techniques for Image and Signal Processing*, 71(2):203–213, December 1998.

[3] L. Baum, T. Petrie, T. Soules, and N. Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of markov cahins. *The Annals of Mathematical Statistics*, 41:164–171, 1970.

[4] M. Brand. An entropic estimator for structure discovery. *Advances in Neural Info. Proc. Systems*, 11:723–729, 1999.

[5] T. Caelli, A. McCabe, and G. Briscoe. Shape tracking and production using hidden markov models. *International Journal of Pattern Recognition and Artificial Intelligence (IJPRAI)*, 15(1):197–221, February 2001.

[6] L. D. Cohen and I. Cohen. Finite-element methods for active contour models and balloons for 2-D and 3-D images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(11):1131–1147, 1993.

[7] C. A. Davatzikos and J. L. Prince. An active contour model for mapping the cortex. *IEEE Transactions on Medical Imaging*, 14(1):65–80, 1995.

[8] R. I. A. Davis, B. C. Lovell, and T. Caelli. Improved estimation of hidden markov model parameters from multiple observation sequences. In *Proceedings of the International Conference on Pattern Recognition (ICPR2002)*, volume 2, pages 168–171, Quebec City, August 2002. IEEE.

[9] D. Geiger, A. Gupta, L. Costa, and J. Vlontzos. Dynamic programming for detecting, tracking, and matching deformable contours. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(3):294–302, 1995.

[10] S. R. Gunn. *Dual Active Contour Models for Image Feature Extraction*. University of Southampton, May 1996. PhD Thesis.

[11] S. E. Levinson, L. R. Rabiner, and M. M. Sondhi. An introduction to the application of the theory of probabilistic functions of a markov process to automatic speech recognition. *The Bell System Technical Journal*, 62(4):1035–1074, 1983.

[12] D. J. C. Mackay. Ensemble learning for hidden markov models. Technical report, University of Cambridge, 1997.

[13] B. McCane and T. Caelli. Diagnostic tools for evaluating hmm components. Technical report, Department of Computer Science, University of Otago, 2001.

[14] L. Rabiner and B. Juang. *Fundamentals of Speech Recognition*. Prentice Hall, New Jersey, 1993.

[15] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, February 1989.

[16] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423, July 1948.

[17] A. Stolke and S. M. Omohundro. Best-first model merging for hidden markov model induction. Technical report, International Computer Science Institute, April 1994. TR-94-003.