BINARY PATTERN RECOGNITION USING MARKOV RANDOM FIELDS AND HMMS

George A. Saon and Abdel Belaïd

Centre de Recheche en Informatique de Nancy CRIN-CNRS,
bât. LORIA, Campus Scientifique, 54506 Vandœuvre-Lès-Nancy, France
E-mail: {saon,abelaid}@loria.fr

ABSTRACT

In this paper we present a stochastic framework for the recognition of binary random patterns which advantageously combine HMMs and Markov random fields (MRFs). The HMM component of the model analyzes the image along one direction, in a specific state observation probability given by the product of causal MRF-like pixel conditional probabilities. Aspects concerning definition, training and recognition via this type of model are developed throughout the paper. Experiments were performed on handwritten digits and words in a small lexicon. For the latter, we report a 89.68% average word recognition rate on the SRTP[1] French postal cheque database (7057 words, 1779 scriptors).

1. INTRODUCTION

Hidden Markov model (HMM) techniques imply sequential pattern processing prior to recognition by performing local observations along a given axis. This step generally contrasts with the 2D nature of the modeled data forcing researchers to enlarge HMM formalism. However, it was proved by Levin in [8] that a direct extension of the dynamic time warping algorithm (DTW), which is the basic mechanism of these models, to the plane, results in an NP-complete problem. By applying a class of constraints to the matching, the complexity can be pulled down to polynomial. A type of models issued from such a simplification are the PHMMs (planar- or pseudo HMMs) [2]. Although these models are easy to implement, the underlying statistical line-independency hypothesis does not always hold true in practice.

We think that a two-dimensional model akin to an image recognition task would be more profitable. Therefore, we have studied the applicability of Markov random fields to binary pattern recognition (BPR). These models naturally overcome the PHMM limitation given that the probabilities of pixels are conditioned by their direct 2-D neighbors. Markov fields have been employed for a long time in statistical mechanics, the application of these models to images being more recent. They are used in image processing and artificial vision in tasks such as segmentation and restoration [6]. We restricted our attention to the study of causal MRFs for two major reasons. First, as stated in [3], one cannot specify arbitrary conditioning neighborhoods for consistency reasons (existence of the joint field probability), whereas there are several theoretical achievements on causal MRFs. On the other hand, recursive training and recognition procedures are more easily applicable on causal fields allowing a natural progression of the joint field mass probability calculus. The concept of causality may have different interpretations since the plane is not provided with a natural order. Two types of causal MRFs are widely used in image processing: the Markov random mesh (MRM) [1] and the unilateral Markov random field also called non-symmetric half-plane Markov chain (NSHP) [10]. Jeng in [7] noted that NSHPs are more appropriate than MRMs when an accurate model for representing two dimensional data is required (MRMs are conditionally independent on 45^o diagonals which diminishes their capability to detect segments having these orientations).

Park and Lee introduced a third-order hidden Markov mesh random field model (HMMRF) and applied it to BPR (handwritten Hangul characters) [9]. By using a vector quantization technique, the cells resulting from a regular decomposition of the input image are encoded into a 2-D sequence of symbols from a finite alphabet. During the decoding phase, a fast *look-ahead* scheme [4] based on marginal MAP is used to find an efficient estimation of the hidden states. The state sequence found is a realization of an underlying stationary Markov random mesh process. In order to avoid the exponential complexity inherent to a complete state decoding, the authors use a simplified version of the 2D EM algorithm (well explained in [5]) called *decision-directed* which was first proposed by Devijver [4]. This algorithm assumes that the lines and the columns are mutually independent which may decrease the modeling accuracy for specific applications as was noticed by the author.

Our approach [11, 12] consists of using NSHP Markov random fields at a pixel observation level without making any hypothesis on the dependency between lines and/or columns. In each state of the HMM, we observe the image columns in a left-to-right fashion. For the current column, the emission probability is computed using state-related NSHP-like conditional pixel distributions. A transition from a state to another implies an optimal change of these distributions in order to maximize the likelihood of the image. Training is based on MLE optimization and mainly consists in estimating the pixel distributions. The estimation is done by performing a maximum likelihood count of pixel configurations (value of the current pixel and of its neighbors) mod-

[1]Service de Recherche Technique de la Poste

ulated by the probability of being in a given state which, at its turn, is expressed using modified *Forward-Backward* functions. The former functions serve also for computing the emission probability during recognition.

2. NON-SYMMETRIC HALF-PLANE HIDDEN MARKOV MODELS

For the following definitions and properties related to MRFs, the reader may refer to [3]. Let $X = \{X_{ij}\}_{(i,j)\in L}$ be a random field defined over a $m \times n$ integer lattice L. X^j stands for the column j of X. Moreover, $P(X_{ij}|X_{kl})$ means the probability of realization x_{ij} of X_{ij} knowing realization x_{kl} of X_{kl}, that is $P(X_{ij} = x_{ij}|X_{kl} = x_{kl})$. Finally, the notation $P(X_{ij}|X_A)$, $A \subset L$, stands for $P(X_{ij}|X_{kl}, (k,l) \in A)$. Since we deal with binarized images, we only consider binary random fields, meaning that random variables take values of $\{0, 1\}$ (0-white pixel, 1-black pixel). According to the previous assumptions, a sample pattern image is naturally one possible realization of a random field. Let us next define the NSHP Markov chain. Consider the following sets:

$$\Sigma_{ij} = \{(k,l) \in L \mid l < j \text{ or } (l = j,\ k < i)\},\ \Theta_{ij} \subset \Sigma_{ij}$$

Σ_{ij} is called the *non-symmetric half-plane* and Θ_{ij} the *support* of pixel $(i,j) \in L$. Both types of sets are illustrated in Figure 1.

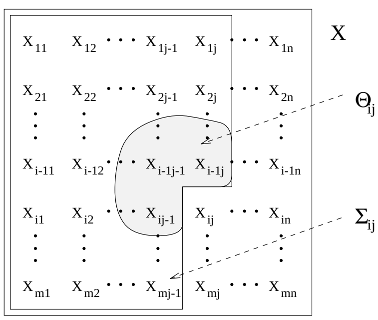

Figure 1: Sets of pixels related to site (i,j).

X is a non-symmetric half-plane Markov chain if and only if:

$$P(X_{ij}|X_{\Sigma_{ij}}) = P(X_{ij}|X_{\Theta_{ij}}), \qquad \forall\ (i,j) \in L \qquad (1)$$

The joint field mass probability $P(X)$ may be computed following the chain decomposition rule of conditional probabilities:

$$P(X) = \prod_{j=1}^{n}\prod_{i=1}^{m} P(X_{ij}|X_{\Sigma_{ij}}) = \prod_{j=1}^{n}\prod_{i=1}^{m} P(X_{ij}|X_{\Theta_{ij}}) \quad (2)$$

Commonly, authors using NSHP Markov chains, choose for all Θ_{ij}'s the same form.

NSHP Markov chains can be implemented by HMMs if we consider the random field realization (pattern image) as an observation sequence of columns. In a specific state of the HMM, observation probability would be given by the column product of pixel conditional probabilities. A transition from one state of the model to another will result in changing the set of probability distributions, and in dynamically modifying feature sensitivity. After the training phase, the model will associate states to particular features within the image areas. The previously mentioned reasons, plus the fact that there are optimal training and recognition procedures, lead us to use HMMs to efficiently implement NSHPs. Figure 2 illustrates the implementation scheme of an NSHP-HMM model.

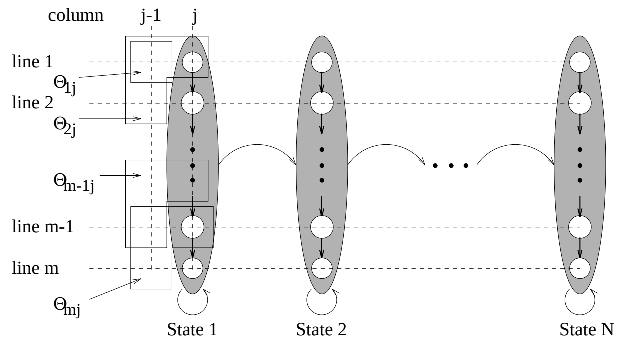

Figure 2: Architecture of an NSHP-HMM model.

Let $Q = q_1 \ldots q_n$ be a stochastic state process associated to the columns of X. The random variables q_j take values in a finite set of states $S = \{s_1, \ldots, s_N\}$. Using equation (2), let us write the pattern likelihood with respect to a model λ ($P(q_1|q_0)$ stands for $P(q_1)$):

$$\begin{aligned} P(X|\lambda) &= \sum_Q P(X,Q|\lambda) = \sum_Q P(X|Q,\lambda)P(Q|\lambda) \\ &= \sum_Q \prod_{j=1}^{n} P(q_j|q_{j-1})P(X^j|X^{j-1}\ldots X^1, q_j, \lambda) \\ &= \sum_Q \prod_{j=1}^{n} P(q_j|q_{j-1}) \prod_{i=1}^{m} P(X_{ij}|X_{\Sigma_{ij}}, q_j, \lambda) \\ &= \sum_Q \prod_{j=1}^{n} P(q_j|q_{j-1}) \prod_{i=1}^{m} P(X_{ij}|X_{\Theta_{ij}}, q_j, \lambda) \end{aligned} \qquad (3)$$

under the assumption that Q is a first order Markov process and that pixel distributions for column j depend only on state q_j. Obviously, (3) bears a strong resemblance to the classical 1D HMM deduction with the difference that we maintain 2D distributions, which we tie to specific states of our HMM. Thus, the difference with an ordinary (discrete) HMM is the giving of the following two elements:

- $\Theta = \{\Theta_{ij}\}_{(i,j)\in L}$, $\Theta_{ij} = \{(i - i_k, j - j_k)|1 \le k \le P, j_k > 0 \text{ or } (j_k = 0, i_k > 0)\} \cap L$, where P represents the number of neighboring pixels per site. Θ is called the neighborhood set and P the *order* of the model.
- $B = \{b_{ik}(x, \mathbf{x})\}_{1\le i\le m, 1\le k\le N}$, $x \in \{0,1\}, \mathbf{x} \in \{0,1\}^P$, $b_{ik}(x,\mathbf{x}) = P(X_{ij} = x|X_{\Theta_{ij}} = \mathbf{x}, q_j = s_k)$, the conditional pixel observation probabilities.

For simplicity, we will denote henceforth an NSHP-HMM by $\lambda = (\Theta, A, B, \pi)$ where A represents the state transition matrix and π the set of initial probabilities. In the following, we show how to estimate the emission probability of a pattern (the image likelihood) and we give some elements concerning training and recognition.

An optimal evaluation of the likelihood $P(X|\lambda)$ is obtained using modified *forward-backward* functions. We will define the *forward* function α (*backward* function β following a dual definition) as being the accumulated field probability until column X^j of X when ending in state s_i, $\alpha_j(i) = P(X^1 X^2 \dots X^j, q_j = s_i|\lambda)$:

$$\begin{aligned} \alpha_1(i) &= \pi_i \prod_{k=1}^{m} b_{ki}(X_{k1}, X_{\Theta_{k1}}), \qquad 1 \le i \le N \\ \alpha_j(i) &= \left[\sum_{l=1}^{N} \alpha_{j-1}(l) a_{li} \right] \prod_{k=1}^{m} b_{ki}(X_{kj}, X_{\Theta_{kj}}),\ j = 2 \dots n \\ P(X|\lambda) &= \sum_{i=1}^{N} \alpha_n(i) \end{aligned} \tag{4}$$

During training, the goal is to determine the parameters (A, B, π) of the model which maximize the product $\prod_{r=1}^{R} P(X^{(r)}|\lambda)$, where $X^{(r)}$ are sample images used to train the model λ. Note that, as in the 1D case, there is no global optimization criterion and direct method. We use the maximum likelihood criterion (MLE) by performing Baum-Welch re-estimation. We will only detail the conditional pixel probability re-estimation:

$$\bar{b}_{il}(x, \mathbf{x}) = \begin{cases} \dfrac{\displaystyle\sum_{r=1}^{R} \frac{1}{P_r} \sum_{\substack{j=1\ s.t.\\ X_{ij}^{(r)} = x\ and \\ X_{\Theta_{ij}}^{(r)} = \mathbf{x}}}^{n_r} \alpha_j^r(l) \beta_j^r(l)}{\displaystyle\sum_{r=1}^{R} \frac{1}{P_r} \sum_{\substack{j=1\ s.t.\\ X_{\Theta_{ij}}^{(r)} = \mathbf{x}}}^{n_r} \alpha_j^r(l) \beta_j^r(l)}, & \text{den} \neq 0 \\ b_{il}(x, \mathbf{x}), & \text{otherwise} \end{cases}$$

$$x \in \{0,1\}, \qquad \mathbf{x} \in \{0,1\}^P, \qquad 1 \le i \le m, \qquad 1 \le l \le N \tag{5}$$

where by $P_r = P(X^{(r)}|\lambda)$, we understand the emission probability of sample $X^{(r)}$ and by n_r its length. Let us take a closer look to equation (5). In fact, pixel probability re-estimation is done by performing an ML count of the number of times that a given pixel configuration is encountered. Note that all samples are supposed to have the same number of lines m which necessitates a height normalization procedure prior to training or recognition.

We chose a *model discriminant* approach by constructing an NSHP-HMM model for each different class. Recognition is performed simply by calculating the pattern likelihood for all models and by labeling the image according to the model which produces the maximum *a posteriori* probability via Bayes decision rule.

3. EXPERIMENTS AND RESULTS

The first set of experiments was conducted on a multi-scriptor database of 562 digits. 337 randomly chosen digit images (60%) served for model training and a distinct set of 225 were used for testing. All images were scaled to $m = 16$ lines and we opt for NSHP-HMM models of order 2 with $N = 10$ states (see Figure 3). Finally, we obtained a 98.22% top 1 and a perfect top 3 digit recognition score.

By far the most relevant experiments were performed on the SRTP database (7057 images, lexicon of 27 words, 1779 scriptors). This BPR task is extremely difficult because of the totally unconstrained writings involved [11]. Images were scanned at 300 dpi from real postal cheques and horizontal and diagonal bars have been removed. The only preprocessing we apply is word-image height normalization.

We randomly chose 5284 images (approximatively 3/4 of the database) for performing word model training. Recognition was done on 1773 distinct images and we report a 89.68% top 1 word recognition rate. Next, we show the initial parameters for each model.

- *State number*: it is proportional to the average word length in pixel columns, $\overline{n}$, after height normalization. In practice, a number of states equal to $\overline{n}/2$ (varying from 11 to 35 for $m = 20$ lines) gave the best recognition results.
- *State transitions*: we allow only transitions to the current or to the next state (strict left-to-right architecture). Initially, transitions are equiprobable, that is $a_{ii} = a_{ii+1} = 0.5,\ 1 \le i \le N-1$.
- *Number of lines*: for computational trainability reasons, we limited this number to $m = 20$. Experiments were carried out with $m = 10$, $m = 15$ and $m = 20$.
- *Model order (number of neighborhood pixels)*: we experimented models of order $P = 0 \dots 4$ corresponding to the neighborhoods depicted in Figure 3.
- *Conditional pixel observation probabilities*: all samples were divided in N vertical bands of equal width. A normalized count of the number of pixel configurations $X_{ij}^{(r)} = x$ and $X_{\Theta_{ij}}^{(r)} = \mathbf{x}$, $\forall x \in \{0,1\}, \mathbf{x} \in \{0,1\}^P$, within each band is performed over all samples $X^{(r)}$.

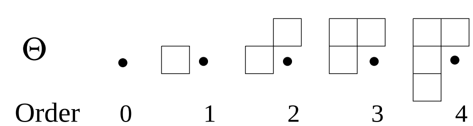

Figure 3: Various neighborhoods considered during testing.

Figure 4 gives us visual feedback on the real learning capabilities of the models. The grey levels code the probability of black pixels, and depend upon the state and the line index of the NSHP-HMM. The word prototypes were obtained using models of order 3 trained with samples of height $m = 30$ (20 iteration steps). One may observe that, despite the huge variability of the patterns at pixel level, the models are able to focus on pixel distributions characterizing specific writing strokes.

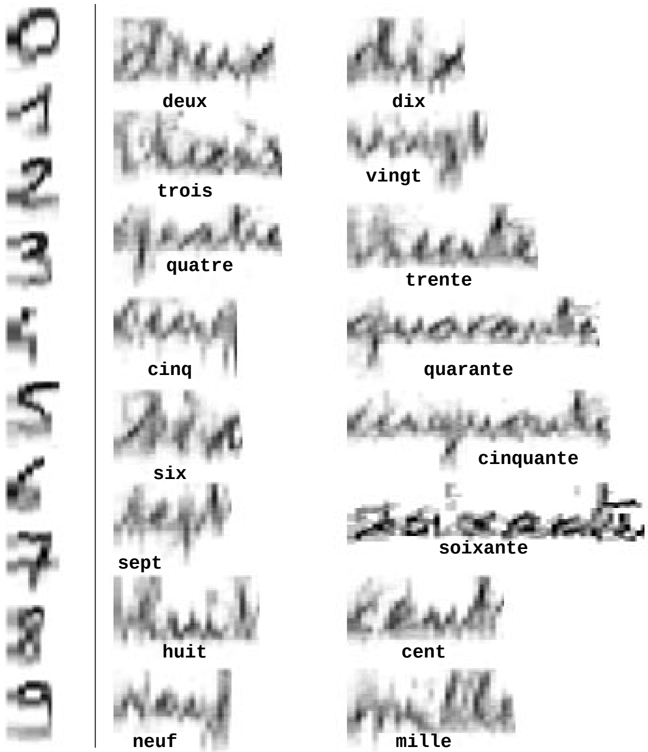

Figure 4: Digit and word prototype synthesis.

4. CONCLUSION

In this paper we have described a new approach to binary pattern recognition which combines causal MRF-like two-dimensional modeling and HMMs. A sample image is viewed as a random field realization which, at its turn, is considered to be an observation sequence of pixel columns. The emission probability of this sequence is calculated using state dependent conditional pixel probabilities. We have seen throughout the article how the estimation of these NSHP-like probabilities is performed by keeping a close relation with the major benefits of HMM formalism (dynamic warping, Baum-Welch re-estimation algorithm, MLE optimization criterion, etc.). The application of these models to handwritten digit and word recognition shows encouraging results and leaves our system open to further improvements. Future development will concern the automatic inference of neighborhoods using the Akaike or Rissanen information criterion suited to two-dimensional data. Another problem which also needs to be addressed is the efficiency of our parameter estimation method MLE compared to methods based, for example, on maximum mutual information.

5. REFERENCES

[1] K. Abend, T. J. Harley, and L. N. Kanal. Classification of Binary Random Patterns. *IEEE Trans. Inform. Theory*, 1(11):538–544, 1965.

[2] O. E. Agazzi and S. Kuo. Hidden Markov Model Based Optical Character Recognition in the Presence of Deterministic Transformation. *Pattern Recognition*, 26(12):1813–1826, February 1993.

[3] H. Derin and P. A. Kelly. Discrete-Index Markov-Type Random Processes. *Proceedings of the IEEE*, 77(10):1485–1510, 1989.

[4] P. A. Devijver. Image Segmentation Using Causal Markov Random Fields Models. *Pattern Recognition*, pages 130–143, 1988.

[5] P. A. Devijver and M. M. Dekesel. Learning the Parameters of a Hidden Markov Random Field Image Model: a Simple Example. *Pattern Recognition Theory and Applications*, pages 140–163, 1987.

[6] F. Heitz and P. Bouthemy. Multimodal Motion Estimation and Segmentation using Markov Random Fields. In *Proceedings of the 10th International Conference on Pattern Recognition (ICPR'90)*, volume 1, pages 378–383, Atlantic City, New Jersey, USA, 16-21 June 1990.

[7] F. C. Jeng and J. W. Woods. On the Relationship of the Markov Mesh to the NSHP Markov Chain. *Pattern Recognition Letters*, 5(4):273–279, 1987.

[8] E. Levin and R. Pieraccini. Dynamic Planar Warping for Optical Character Recognition. In *IEEE-ICASSP*, volume III, pages 149–152. IEEE, 1992.

[9] H. S. Park and S. W. Lee. An HMMRF-Based Statistical Approach for Off-line Handwritten Character Recognition. In *IEEE Proceedings of ICPR'96*, volume 2, pages 320–324, 1996.

[10] D. Preuss. Two-Dimensional Facsimile Source Coding Based on a Markov Model. *NTZ 28*, 5(4):358–363, 1975.

[11] G. Saon and A. Belaïd. Recognition of Unconstrained Handwritten Words Using Markov Random Fields and HMMs. In *Fifth International Workshop on Frontiers in Handwriting Recognition (IWFHR5)*, pages 429–432, University of Essex, England, September 1996.

[12] G. Saon and A. Belaïd. High Performance Unconstrained Word Recognition System Combining HMMs and Markov Random Fields. *International Journal of Pattern Recognition and Artificial Intelligence Special Issue on Automatic Bankcheck Processing*, 1997. to appear.