Champs aléatoires de Pickard et modélisation d'images digitales

Digital image modeling using Pickard random fields

Pierre A. DEVIJVER

Philips Research Laboratory, avenue Van Becelaere 2, Bte 8, B-1170 BRUXELLES, BELGIQUE.

Ingénieur civil des télécommunications, Bruxelles, 1962, Docteur Ingénieur, Paris-VI, 1977. Au Philips Research Laboratory Brussels (PRLB) depuis 1969, il a conduit de nombreux travaux de recherche dans le domaine de la reconnaissance des formes. Actuellement, il est responsable de la recherche en analyse et interprétation d'images. Ses travaux les plus récents concernent la modélisation spatio-temporelle des séquences d'images fondée sur l'utilisation de champs aléatoires markoviens multidimensionnels.

Depuis 1980, Pierre A. Devijver a assumé diverses fonctions au sein du comité exécutif de l'IAPR (International Association for Pattern Recognition) dont il fut le président de 1986 à 1988.

Michel DEKESEL

MBLE SA, 82, rue des Deux Gares, 1070 BRUXELLES, BELGIQUE.

École Technique Supérieure en Électronique et Informatique, Bruxelles, 1972. Il a collaboré à des projets de recherche en traitement et reconnaissance des images au Laboratoire de Recherche Philips de 1973 à 1988. Il a été responsable du développement d'un logiciel de traitement interactif d'images implanté dans la plupart des laboratoires Philips européens. Actuellement, au département BEDT de la société MBLE il participe à un projet de codage et compression d'images dans le cadre de la « Belgian Broadband Association ».

RÉSUMÉ

Dans cet article, nous développons un modèle d'image qui fait appel aux champs aléatoires markoviens de Pickard dans le but de modéliser des notions contextuelles aussi vagues et imprécises que « l'uniformité d'une région » ou « la continuité du bord d'un objet ». Nous décrivons une méthode d'estimation par maximum de vraisemblance *a posteriori* obtenue par une généralisation simple d'une méthode largement utilisée dans le contexte unidimensionel de la reconnaissance de la parole. Nous développons deux méthodes d'estimation non supervisée des paramètres du modèle et nous montrons au moyen de plusieurs exemples que notre technique permet de traiter avec succès des problèmes de restauration et de segmentation d'images digitales à niveaux de gris.

MOTS CLÉS

Chaînes de Markov, champs de Markov, réseaux de Markov, champs de Pickard, modélisation markovienne de la parole, modélisation markovienne de l'image, restauration d'image, segmentation d'image.

SUMMARY

This paper outlines a modeling technique for digital images which relies on Markov random fields proposed by Pickard for the purpose of representing fuzzy contextual concepts such as "the uniformity of a region" or "the continuity of a contour". We develop a maximum likelihood estimation technique which is a straightforward generalization of an approach which is used quite extensively in speech recognition circles. Next, we outline two nonsupervised parameter estimation techniques which enable us to infer the model parameters from actual imagery data. We offer a number of practical examples providing evidence that our approach is well suited to handle problems of image restauration and/or segmentation.

KEY WORDS

Markov chains, Markov fields, Markov meshes, Pickard random field, Markov models for speech, Markov image models, image restauration, image segmentation.

1. Introduction

Au cours des dernières années, on a vu se développer un intérêt considérable pour les méthodes de restauration et de segmentation d'images fondées sur la technique d'estimation par maximum de vraisemblance *a posteriori* dans des modèles d'image doublement stochastiques dont une des composantes, au moins, est de nature markovienne [8, 15, 17, 20, 22, 23, 26-28, 31, 54]. Dans ces modèles, la finalité de la composante markovienne est de modéliser la notion fort imprécise de contexte spatial qui se traduit par des expressions telles que « l'uniformité d'une région » ou « la continuité du bord d'un objet ». La nature stochastique de ces modélisations du contexte contraste avec celle, déterministe, qui sous-tend, par exemple, l'algorithme Phagocyte de Brice et Fennema [10] et l'algorithme « Split and Merge » de Horowitz et Pavlidis [33, 45]. Ces notions contextuelles ne donnant pas lieu à des mesures quantitatives par un capteur optique, nous dirons que la composante markovienne est *non observable.*

Les méthodes qui ont été proposées à ce jour exploitent divers modèles d'image et divers critères d'optimalité, et couvrent une large gamme de l'échelle de complexité. Ainsi, à titre d'exemple, la famille des composantes markoviennes comprend-elle les champs de Markov [38] dans [8] et [26], les réseaux de Markov [2, 36] dans [15], [17], [20] et [39], et les champs de Pickard [46, 47] dans [22], [23] et [31]. Il est incontestable que la technique la plus efficace, s'appuyant sur le traitement théorique le plus exhaustif, soit à mettre au crédit de Geman et Geman, [26]. Toutefois, [26], pas plus que les autres travaux évoqués ci-dessus – exception faite de [22] et [23] – n'apporte de réponse au problème crucial de l'estimation non-supervisée des paramètres du modèle d'image doublement stochastique à composante markovienne ([1]). Le lecteur trouvera dans un travail récent de Geman, Geman et Graffigne [27] une formulation très générale du problème, une discussion de sa complexité, et le souhait de voir mettre au point une solution simple fondée sur des principes statistiques solides.

Une situation totalement différente prévaut dans le domaine de la reconnaissance de la parole où un modèle doublement stochastique, dont la composante markovienne est une chaîne de Markov non observable, a été utilisé avec un succès remarquable par de nombreuses équipes de recherche au cours de la dernière décennie. Dans le domaine uni-dimensionnel, diverses techniques de reconnaissance (classement) et d'apprentissage (estimation de paramètres) ont été développées et ont fait l'objet d'un très grand nombre de publications [1, 3-5, 9, 13, 16, 18, 25, 30, 34, 35, 41, 44, 49]. (Le lecteur peu familiarisé avec les modèles markoviens de la parole trouvera dans [42] et [48] un exposé introductif fort complet.)

En ce qui concerne l'apprentissage, c'est-à-dire l'estimation non supervisée des paramètres du modèle, on pourrait imaginer que les idées et les méthodes qui ont été mises en œuvre au bénéfice de la parole aient pu se prêter à une transposition au bénéfice de l'image. Nous n'avons pas connaissance qu'il en ait été ainsi. Ceci peut s'expliquer par un certain nombre de raisons dont la plus évidente réside dans une différence essentielle entre les modèles postulés : d'une part, la chaîne de Markov est un processus stochastique *causal* tandis que le champ markovien est, en général, un processus *non causal.* Cette constatation reflète bien l'existence d'un ordonnancement temporel et naturel des échantillons du signal de parole, et l'absence d'ordonnancement naturel des pixels qui composent une image. A cette différence intrinsèque, s'ajoute la difficulté liée au passage d'un modèle unidimensionnel de parole au modèle bi-dimensionnel d'image. Ainsi, si la causalité uni-dimensionnelle est synonyme de récursivité, un modèle bi-dimensionnel, fût-il même causal (*ex. le réseau markovien*), ne se prête à un traitement récursif qu'au prix d'hypothèses simplificatrices contraignantes. La conjugaison de ces disparités suffirait à elle seule à justifier l'absence d'approche multidisciplinaire.

Il existe cependant une catégorie tout à fait remarquable de modèles d'image doublement stochastiques dont la composante markovienne jouit simultanément des propriétés de causalité et de non-causalité. Il s'agit des modèles qui font appel aux champs de Pickard [46, 47], une sous-classe de la classe des champs de Markov, et une sous-classe de la classe des réseaux de Markov. Qui plus est, on retrouve dans les champs bi-dimensionnels de Pickard, des structures de chaînes uni-dimensionnelles de Markov. Par exemple, nous verrons à la section 2.2 que les lignes (colonnes) de tout ensemble de k colonnes (lignes) consécutives d'un champ de Pickard forment une chaîne de Markov vectorielle (de dimension k), $k = 1, 2, \ldots$ Ainsi, les champs de Pickard bénéficient-ils de certaines propriétés exploitées par les algorithmes récursifs de traitement de parole, tout en fournissant, au prix de quelques hypothèses simplificatrices, un modèle d'image tout à fait légitime.

L'idée d'exploiter les structures de chaîne de Markov apparaissant dans les champs de Pickard dans le but de développer une technique de reconnaissance par maximum de vraisemblance *a posteriori* applicable à l'image est à mettre au crédit de Haslett [31]. Ce faisant, cet auteur fut naturellement amené à redécouvrir un algorithme récursif bien connu des spécialistes de la reconnaissance de la parole.

Dans ce qui suit, nous ne ferons que prolonger la démarche de Haslett, en nous intéressant plus particulièrement au problème d'apprentissage que ce dernier n'avait pas abordé. Il va quasi de soi que notre démarche s'inspirera fortement des techniques d'apprentissage utilisées en reconnaissance de la parole. Ainsi, nous développerons à la section 4.2 un algorithme d'identification de mélange de distributions de type EM (de la terminologie anglo-saxonne, E pour

([1]) Il y a lieu de faire une distinction claire entre les problèmes d'apprentissage qui se posent dans les modèles *simplement* et *doublement* stochastiques. Le prototype du cas « simplement » stochastique consiste à modéliser une image de texture par un champ markovien. Dans ce cas, le champ est directement observable, et les paramètres qui le caractérisent peuvent être estimés par des techniques classiques de maximisation de la pseudo-vraisemblance, [7]. Cette méthode a été utilisée avec succès par Cross et Jain [14], et Geman, Geman et Graffigne [27]. Par contre, lorsque la composante markovienne ne peut être observée, comme dans le cas qui nous occupe, le problème devient considérablement plus ardu.

Expectation, M pour Maximization [50, 19] qui prend en compte la structure markovienne particulière de notre modèle et qui est une généralisation de la méthode développée par Baum *et al.* [4, 5, 42]. Cet algorithme de type EM nous fournira le point de départ de la formulation d'un algorithme, plus économique en temps calcul, de type DD (à nouveau, de la terminologie anglo-saxonne, D pour Decision, D pour Directed) qui, à notre connaissance n'a pas d'équivalent en reconnaissance de parole. Nous montrerons, à la section 5, que notre technique permet de traiter avec succès des problèmes de restauration et de segmentation d'images digitales à niveaux de gris.

La formulation de nos algorithmes repose, dans une large mesure, sur l'algorithme *Forward-Backward* ([2]) de Baum [4, 5]. Nous en rappellerons les principes à la section 3, où nous montrerons également que l'algorithme de Haslett [31] peut se ramener à deux exécutions « orthogonales » de l'algorithme *Forward-Backward* (une fois le long des lignes et une fois le long des colonnes de l'image). Auparavant, à la section 2, nous introduirons brièvement les modèles doublement stochastiques de la parole et de l'image.

2. Modèles doublement stochastiques

2.1. Le modèle uni-dimensionnel

Cette section propose un bref rappel des propriétés des modèles doublement stochastiques dont la composante non observable est une chaîne de Markov, rappelle certaines propriétés qui seront utiles dans la suite, et introduit les notations nécessaires. Après avoir défini le modèle abstrait, nous montrerons qu'il correspond bien à divers problèmes qui se posent en reconnaissance des formes et, en particulier, au problème de reconnaissance de la parole.

Soit ω une variable aléatoire dont la loi de distribution est une chaîne de Markov homogène, du premier ordre, et dont l'espace d'état (discret) est noté $\Psi \doteq \{\psi_1, \ldots, \psi_\vartheta\}$. Nous écrirons $\omega^\tau = \psi_j$ pour indiquer que le processus est dans l'état ψ_j au temps τ. La chaîne de Markov est caractérisée par une distribution initiale $P_j = P(\omega^1 = \psi_j)$, $j = 1, \ldots, \vartheta$, et une matrice de *probabilités de transition*, que nous supposons stationnaires, $P_{jk} = P(\omega^{\tau+1} = \psi_k / \omega^\tau = \psi_j)$, $\tau \geqq 1$, et $j, k \in \{1, \ldots, \vartheta\}$. Rappelons qu'une chaîne de Markov du premier ordre est définie par la propriété que pour tout $\tau > 1$, $P(\omega^\tau/\omega^1, \ldots, \omega^{\tau-1}) = P(\omega^\tau/\omega^{\tau-1})$. Il s'ensuit que la probabilité d'une réalisation finie arbitraire $\{\psi_{i_\tau}\}_{\tau=1}^{T}$ peut être factorisée sous la forme

$$(1)\quad P(\psi_{i_1}, \ldots, \psi_{i_T}) = P(\omega^1 = \psi_{i_1}) \prod_{\tau=2}^{T} P(\omega^\tau = \psi_{i_\tau} / \omega^{\tau-1} = \psi_{i_{\tau-1}}) = P_{i_1} \prod_{\tau=2}^{T} P_{i_{\tau-1} i_\tau},$$

qui traduit bien la structure *causale* de la chaîne. Dans ce qui suit, nous supposerons avoir affaire à une chaîne *régulière*, c'est-à-dire sans état transitoire, et ne comportant qu'un ensemble ergodique et une seule classe cyclique.

Il sera utile dans la suite de garder présent à l'esprit le fait qu'une chaîne de Markov considérée en sens inverse, c'est-à-dire selon les valeurs décroissantes de τ, est aussi un processus markovien. Il ne s'agit toutefois pas nécessairement d'une chaîne, les probabilités de transition régressives étant, en général, dépendantes de τ. Parmi les chaînes qui jouissent de la propriété de conserver la nature de chaîne en sens inverse, nous nous intéresserons plus particulièrement aux chaînes dites *réversibles*, qui sont la réalisation d'un même processus stochastique indépendamment du sens d'observation. D'une manière formelle, soit $[\pi_1, \ldots, \pi_\vartheta]$ la distribution stationnaire d'une chaîne de Markov régulière; la chaîne est dite réversible si, et seulement si,

$$P(\omega^\tau = \psi_i \cap \omega^{\tau+1} = \psi_j) = P(\omega^\tau = \psi_j \cap \omega^{\tau+1} = \psi_i),$$

uniformément en τ, ou, de manière équivalente, $\pi_i P_{ij} = \pi_j P_{ji}$. On peut montrer [37] que :

(i) toute chaîne de Markov *binaire* (à deux états) est réversible;

(ii) toute chaîne de Markov ergodique dont la matrice des probabilités de transition est symétrique est réversible.

Nous supposerons que la chaîne de Markov n'est pas accessible à l'observation. Par contre, ce qui peut être observé est une variable aléatoire X qui dépend de ω par l'intermédiaire de ϑ distributions $p_j(X) = p(X/\psi_j)$, $j = 1, \ldots, \vartheta$. De plus, étant donné une séquence $\mathbf{X}_1^T \doteq \{X^1, \ldots, X^T\}$ d'observations de la variable aléatoire X — où X^τ est l'observation au temps τ — et la séquence $\{\omega^1, \ldots, \omega^T\}$ correspondante, nous émettrons une hypothèse d'indépendance mutuelle des $X^1, \ldots, X^T$ conditionnellement à $\omega^1, \ldots, \omega^T$. En d'autres mots, nous supposerons que, étant donné ω^τ, X^τ est stochastiquement indépendant de $\omega^{\tau'}$ et $X^{\tau'}$ pour tout $\tau' \neq \tau$. Il s'ensuit une seconde factorisation :

$$(2)\quad p(X^1, \ldots, X^T/\omega^1, \ldots, \omega^T) = \prod_{\tau=1}^{T} p(X^\tau/\omega^\tau).$$

En combinant les factorisations en (1) et (2) et en faisant appel au théorème de la probabilité totale, on peut montrer aisément que la vraisemblance $\mathscr{L}$ de la séquence d'observations $X^1, \ldots, X^T$ s'exprime comme suit

$$(3)\quad \mathscr{L} = p(X^1, \ldots, X^T) = \sum_{\omega^1, \ldots, \omega^T = \psi_1}^{\Psi} P(\omega^1)\, p(X^1/\omega^1) \times \prod_{\tau=1}^{T-1} P(\omega^{\tau+1}/\omega^\tau)\, p(X^{\tau+1}/\omega^{\tau+1}) = \sum_{i_1, \ldots, i_T = 1} P_{i_1} p_{i_1}(X^1) \times \prod_{\tau=1}^{T-1} P_{i_\tau i_{\tau+1}} p_{i_{\tau+1}}(X^{\tau+1}).$$

([2]) Dans le cas de cet algorithme devenu aujourd'hui un « classique » de la reconnaissance de la parole, nous nous permettrons de conserver la terminologie anglo-saxonne.

Dans ce qui suit, nous serons confrontés en permanence au problème de calculer des expressions semblables. Remarquons toutefois que le calcul de $\mathscr{L}$ en (3) requiert $2T\vartheta^T$ multiplications. Dans la plupart des applications réelles – ainsi, par exemple, en reconnaissance de la parole où ϑ et T sont de l'ordre de quelques dizaines – un recours brutal à la formule (3) serait parfaitement hors de question pour des raisons de temps calcul. Par bonheur, la causalité nous permettra de *linéariser* la complexité du calcul par le biais de la récursivité. Pour ce faire, nous ferons essentiellement appel aux travaux de Baum *et al.*, bien que ce problème ait été investigué par de nombreux autres auteurs [3-5, 16, 18, 25, 30].

A titre d'illustration du modèle ainsi défini, considérons brièvement le cas d'une chaîne de Markov a six états, $\psi_1, \ldots, \psi_6$, équi-probables initialement ($P_i = 1/6$, $i = 1, \ldots, 6$). Supposons de plus que la matrice des probabilités de transition soit définie par

$$\begin{aligned} P_{ij} = p > 0 \quad & \text{si} \quad j = i, \\ q > 0 \quad & \text{si} \quad j = i+1 \bmod 6, \\ r > 0 \quad & \text{si} \quad j = i+2 \bmod 6, \\ 0 \quad & \text{sinon,} \end{aligned}$$

avec $p+q+r=1$. Cette matrice s'écrit

$$[P_{ij}] = \begin{bmatrix} p & q & r & & & \\ & p & q & r & & \\ & & p & q & r & \\ & & & p & q & r \\ r & & & & p & q \\ q & r & & & & p \end{bmatrix}$$

(où les éléments non apparents sont nuls).

Il y a diverses formes de représentation pour le modèle doublement stochastique que nous venons de définir. Ici, nous adopterons la représentation en graphe des figures 1-3. Chaque nœud du graphe correspond à un état distinct à un instant donné; ici $\tau_1 \leqq \tau \leqq \tau_{10}$. Chaque branche représente une transition vers un nouvel état à l'instant suivant. On peut voir à la figure 1 une séquence d'états issue de l'état initial $\omega^{\tau_1} = \psi_3$ et se terminant à l'état final $\omega^{\tau_{10}} = \psi_3$. A chaque séquence d'états possible correspond un *chemin* dans le graphe et vice versa. Remarquons que la présence de zéros dans la matrice des probabilités de transition réduit le nombre de chemins possibles au travers du graphe.

Nous avons supposé de plus qu'à chaque instant τ, le système engendre, ou émet, un signal observable X^τ avec une probabilité $p(X^\tau/\omega^\tau)$ qui ne dépend que de l'état ω^τ qui prévaut au temps τ. En vertu de la propriété de Markov et de notre hypothèse d'indépendance conditionnelle, la probabilité conjointe de la séquence d'observations $X^{\tau_1}, \ldots, X^{\tau_{10}}$ et de la séquence d'états $\omega^{\tau_1} = \psi_3, \ldots, \omega^{\tau_{10}} = \psi_3$, illustrées à la figure 1, est le produit des probabilités initiale, de transition et conditionnelles rencontrées le long du chemin choisi dans le graphe. Ce produit est un des termes de la somme apparaissant dans l'équation (3).

Arrêtons-nous un court instant à l'interprétation physique du modèle abstrait que nous venons de définir. Dans le cas de la reconnaissance de la parole, les états ω^τ pourraient servir à caractériser la *classe phonétique* à laquelle appartient un segment du signal de parole. Les observations X^τ pourraient représenter le vecteur de coefficients de prédiction linéaire (ou de coefficients de Fourier, ou encore de coefficients cepstraux) du segment correspondant, détecté par un capteur acoustique. En reconnaissance de caractères, l'interprétation la plus naturelle de la notion d'état s'identifie avec celle de l'identité du caractère à reconnaître *i. e.*, son rang dans l'alphabet, tandis que l'observation pourrait être un vecteur de coefficients de Fourier du contour d'un « objet » détecté dans une image acquise par un capteur optique.

En tout état de cause, les variables X de notre modèle sont supposées *observables*. Les variables d'état ω ne le sont pas. Comme nous le verrons à la section suivante, le premier problème qui nous occupera sera celui de déterminer à chaque instant l'état le plus probable compte tenu d'une séquence d'observations donnée et de la connaissance supposée des paramètres du modèle. Ce problème, et le modèle qui lui est associé, justifient l'utilisation courante de l'expression anglo-saxonne de « *hidden* Markov chain model ».

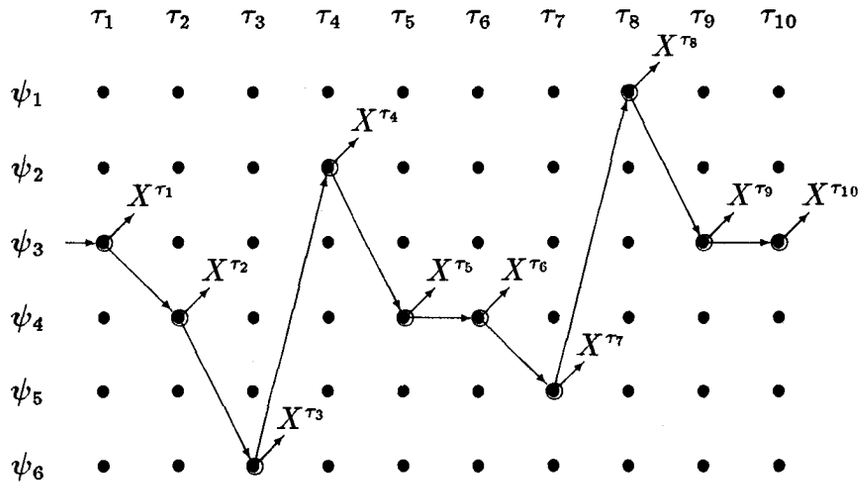


Fig. 1. – Un chemin arbitraire dans la représentation en graphe d'un modèle doublement stochastique dont la composante markovienne est une chaîne à six états.

Avant de quitter le domaine uni-dimensionnel, il nous paraît utile de souligner le fait que certaines propriétés des chaînes de Markov se retrouvent, d'une manière quasi immédiate, dans la distribution conjointe de la variable (ω, X). Ainsi, est-il bien connu que, conditionnellement à ω^τ, les séquences $\{\omega^t\}_1^{\tau-1}$ et $\{\omega^t\}_{\tau+1}^{T}$ sont mutuellement indépendantes. Par analogie, on peut montrer aisément que, sous les hypothèses formulées ci-dessus,

$$(4) \qquad P(X^{\tau+1}/\omega^\tau = \psi_j,\ \bar{X}_1^\tau) = P(X^{\tau+1}/\omega^\tau = \psi_j),$$

et

$$(5) \qquad P(\bar{X}_{\tau+1}^T/\omega^\tau = \psi_j,\ X^\tau) = P(\bar{X}_{\tau+1}^T/\omega^\tau = \psi_j).$$

D'une manière plus générale,

$$(6) \quad P(\bar{X}_1^\tau,\ \bar{X}_{\tau+1}^T/\omega^\tau = \psi_j) = P(\bar{X}_1^\tau/\omega^\tau = \psi_j)\, P(\bar{X}_{\tau+1}^T/\omega^\tau = \psi_j).$$

2.2. Le modèle bi-dimensionel

Nous nous tournons à présent vers les *champs aléatoires* définis sur un sous-ensemble fini et rectangulaire du treillis bi-dimensionnel des entiers (m, n) positifs : $V_{M,N} \doteq \{(m, n) : 1 \leqq m \leqq M,\ 1 \leqq n \leqq N\}$. Il va sans dire que cette représentation traduit notre intérêt pour les images digitales, et que nous ne ferons pas de distinction entre un site (m, n) du treillis et le pixel à l'intersection de la ligne d'indice m et de la colonne d'indice n dans une image digitale de taille $M \times N$. En accord avec la pratique la plus communément admise en traitement d'image, nous supposerons que le premier axe est orienté vers le bas et le second vers la droite, de telle sorte que le pixel (1, 1) soit situé au coin supérieur gauche de l'image. Nous supposons que chaque site (m, n) du treillis est susceptible de se trouver dans un état donné [3] $\lambda_{m,n} \in \Psi$ et que chaque pixel donne lieu à une observation $X_{m,n}$ de certaines caractéristiques locales de l'image.

A ce stade, l'interprétation physique de notre modèle est immédiate : ainsi, un état binaire peut-il indiquer la présence ou l'absence du bord d'un objet dans l'image au pixel considéré; un état multi-valué pourrait-il indiquer la région de l'image à laquelle le pixel appartient. L'observation $X_{m,n}$ pourrait être une simple mesure de niveau de gris local, un vecteur de mesures de réflectances multi-spectrales ou, dans un modèle plus élaboré, un vecteur de descripteurs de texture.

La généralisation, à deux dimensions, de la propriété de Markov est un problème dont la difficulté, comme nous l'avons déjà signalé, trouve son origine dans l'absence d'un ordonnancement naturel des pixels d'une image. La question a été largement débattue par de nombreux auteurs, *e. g.*, [6, 7, 32, 37, 40]. Nous ne ferons ici que rappeler les principaux résultats de ces travaux de manière à situer le champ aléatoire de Pickard au sein des autres familles connues de champs aléatoires markoviens.

La formulation la plus naturelle de la propriété markovienne à deux dimensions, celle qui est la plus généralement admise, et qui engendre ce que nous appelerons *les champs de Markov*, est la suivante :

$$(7) \quad P(\lambda_{m,n}/\lambda_{k,l} : (k, l) \neq (m, n)) = P\left(\lambda_{m,n} \middle/ \begin{matrix} \lambda_{m-1,n-1} & \lambda_{m-1,n} & \lambda_{m-1,n+1} \\ \lambda_{m,n-1} & & \lambda_{m,n+1} \\ \lambda_{m+1,n-1} & \lambda_{m+1,n} & \lambda_{m+1,n+1} \end{matrix}\right),$$

(nous négligeons les ajustements nécessaires aux limites de l'image). On peut voir directement que la propriété (7) est *non causale.*

D'autre part, Kanal *et al.* ont proposé une propriété de « markovianité » qui reproduit partiellement, en terme de sites dans le treillis, les notions de « passé », « présent » et « futur », [2, 36]. Leur définition de la propriété qui engendre un *réseau de Markov du troisième ordre* s'écrit

$$(8) \quad P(\lambda_{m,n}/\lambda_{k,l} : k < m \text{ ou } l < n) = P\left(\lambda_{m,n} \middle/ \begin{matrix} \lambda_{m-1,n-1} & \lambda_{m-1,n} \\ \lambda_{m,n-1} & \end{matrix}\right).$$

Il va de soi que nous avons à faire ici à une propriété *causale.*

Il importe de souligner que (8) implique (7), mais que la réciproque n'est pas vraie en général. En fait, comme le note Besag [7], la famille des réseaux de Markov est une sous-classe dégénérée de la famille des champs de Markov. Toutefois, l'expérience a montré que cette sous-classe reste suffisamment générale pour modéliser une grande variété d'images digitales.

Le champ aléatoire de Pickard [46, 47] jouit de la propriété d'être simultanément un champ de Markov et un réseau de Markov. Par souci de simplicité, nous nous limiterons ici aux champs binaires (à deux états possibles), sans que cette restriction enlève quoi que ce soit à la généralité des résultats énoncés.

Le champ de Pickard est un champ aléatoire homogène, c'est-à-dire dont les propriétés sont indépendantes de la localisation dans le treillis, qui est engendré par la distribution conjointe des états aux sommets d'une cellule unitaire

$$\begin{pmatrix} (m-1, n-1) & (m-1, n) \\ (m, n-1) & (m, n) \end{pmatrix}.$$

Cette distribution est invariante pour les transformations symétriques du carré, et satisfait à la condition

$$(9)\ P\left(\lambda_{m-1,n} \middle/ \begin{matrix} \lambda_{m-1,n-1} \\ \lambda_{m,n-1} \end{matrix}\right) = P(\lambda_{m-1,n}/\lambda_{m-1,n-1}).$$

Pickard a montré [46] que, dans le cas binaire, cette distribution est complètement définie par trois paramètres seulement [4] à savoir,

$$(10) \quad \begin{cases} \theta = P(\lambda_{m,n} = 1), \qquad a = P(\lambda_{m,n} = 1/\lambda_{m,n-1} = 1), \\ d = P\left(\lambda_{m,n} = 1 \middle/ \begin{matrix} \lambda_{m-1,n-1} = 1 & \lambda_{m-1,n} = 1 \\ \lambda_{m,n-1} = 1 & \end{matrix}\right). \end{cases}$$

[3] Par soucis de clarté, nous adoptons des symboles différents pour les variables aléatoires d'état dans les processus uni- et bi-dimensionnels — ceci ne modifie toutefois en rien l'interprétation sémantique de la notion d'état — tout en gardant la même notation pour les deux espaces d'états possibles.

[4] Cette assertion peut être interprétée comme un cas particulier du théorème de Hammersley-Clifford [7] pour un champ de Markov isotrope satisfaisant (9).

[On remarquera que (10) traduit bien la *causalité* du champ aléatoire de Pickard.] Toutefois, la détermination de l'espace des paramètres admissibles — un sous-ensemble fort irrégulier du cube unité — s'avère être un problème particulièrement compliqué.

Posons

$$b = \mathrm{P}(\lambda_{m,n} = 1/\lambda_{m,n-1} = 0).$$

Il vient $b = \theta(1-a)/(1-\theta)$. Du point de vue des méthodes de reconnaissance et d'apprentissage que nous aborderons sous peu, la caractéristique des champs de Pickard qui motive tout l'intérêt que nous leur portons peut s'énoncer comme suit : *Chaque ligne (et colonne) d'un champ de Pickard constitue un segment d'une chaîne de Markov homogène et réversible et dont la matrice des probabilités de transition (stationnaires) s'écrit* $\begin{pmatrix} a & 1-a \\ b & 1-b \end{pmatrix}$. Il s'ensuit immédiatement que

$$\mathrm{P}(\lambda_{m,n}/\lambda_{m,n'};\ n' \neq n) = \mathrm{P}(\lambda_{m,n}/\lambda_{m,n\pm1}),$$

où nous utilisons $(m, n\pm1)$ comme abbréviation pour $\{(m, n-1), (m, n+1)\}$. La même propriété est évidemment vérifiée le long des colonnes. De plus, il s'avère que

$$(11)\quad \mathrm{P}(\lambda_{m,n}/\lambda_{k,l} : k=m \text{ ou } l=n,\ (k,l) \neq (m,n)) = \mathrm{P}(\lambda_{m,n}/\lambda_{m,n\pm1},\ \lambda_{m\pm1,n}).$$

Nous ferons un large usage de ces propriétés dans la suite.

Il n'est pas sans intérêt de s'arrêter brièvement à la structure de corrélation des divers champs markoviens dont nous venons de parler. Rappelons tout d'abord que dans le cas uni-dimensionnel, la corrélation $\rho_{\tau,\tau+t}$ entre les états aux temps τ et $\tau+t$ d'une chaîne de Markov décroît géométriquement en fonction de t. Que ceci réponde bien, ou non, au phénomène physique que l'on cherche à modéliser est une question qui ne peut trouver de réponse que dans le contexte de l'application considérée. (Dans le cas de la parole, certains ont cherché à s'affranchir de cette « contrainte » intrinsèque du modèle, [41].)

Une situation bien différente peut prévaloir dans le cas bi-dimensionnel. On sait, en effet, qu'il existe une équivalence parfaite entre l'ensemble des champs markoviens — au sens de (7) — et les distributions de Gibbs de la mécanique statistique qui décrivent la distribution de probabilité, dans l'espace des phases, d'un système physique en équilibre thermique avec son environnement. Or, il est bien connu qu'en dessous de la température de Curie, apparaissent des configurations régulières du système physique, par exemple l'alignement des spins atomiques dans un aimant, qui traduisent une corrélation positive entre les états de sites séparés par une distance arbitraire. Cette forme de *régularité à grande échelle* qui caractérise certains champs markoviens, n'est évidemment pas souhaitable dans le cas d'un modèle d'image, car elle équivaut à attribuer une forte probabilité *a priori* à une image — supposée infinie pour les besoins du raisonnement — constituée de régions uniformes *d'étendues également infinies*, ce qui aurait pour effet de contrecarrer la restauration des détails d'étendue faible.

Les réseaux markoviens, par contre, ne manifestent pas cette caractéristique défavorable. En ce qui concerne les champs de Pickard, la situation est parfaitement claire, et la structure de corrélation qui les caractérise particulièrement simple. En effet, si $\rho_{k,l}$ est la corrélation entre $\lambda_{m,n}$ et $\lambda_{m+k,n+l}$, on peut montrer que $\rho_{k,l} = (a-b)^{|k|+|l|}$, où a et b sont les paramètres définis ci-dessus [46]. Cette observation confirme bien l'absence de régularité à grande échelle.

Comme à la section précédente, nous supposerons que la variable aléatoire observable X dépend de l'état local par l'intermédiaire de ϑ distributions conditionnelles $p(\mathrm{X}/\psi_j)$, $j=1, \ldots, \vartheta$. En d'autres mots, nous supposerons que la distribution de $\mathrm{X}_{m,n}$ est $p_j(\mathrm{X}_{m,n})$ lorsque $\lambda_{m,n} = \psi_j$. De même, nous supposons que, conditionnellement à $\lambda_{m,n}$, $\mathrm{X}_{m,n}$ est stochastiquement indépendant de $\lambda_{k,l}$ et $\mathrm{X}_{k,l}$ pour tout $(k,l) \neq (m,n)$. Dans ces conditions, il est aisé de démontrer que les propriétés d'indépendance (4)-(6) peuvent être généralisées dans le cas bi-dimensionnel. En posant $\bar{\mathrm{X}}_{m,1}^{m,n-1} \doteq \{\mathrm{X}_{m,1}, \ldots, \mathrm{X}_{m,n-1}\}$ on peut montrer aisément que

$$(12)\quad p(\mathrm{X}_{k,l} : k=m \text{ ou } l=n,\ (k,l) \neq (m,n)/\lambda_{m,n}) = p(\bar{\mathrm{X}}_{m,1}^{m,n-1}/\lambda_{m,n})\, p(\bar{\mathrm{X}}_{m,n+1}^{m,\mathrm{N}}/\lambda_{m,n}) \times p(\bar{\mathrm{X}}_{1,n}^{m-1,n}/\lambda_{m,n})\, p(\bar{\mathrm{X}}_{m+1,n}^{\mathrm{M},n}/\lambda_{m,n}).$$

C'est sur cette observation que se fonde le travail de Haslett [31] que nous évoquerons à la section 3.2.

3. Algorithmes de reconnaissance

La question qui va nous occuper présentement est celle d'obtenir des algorithmes efficaces permettant de calculer les probabilités des états, étant donné les observations et les paramètres du modèle sous-jacent. Du point de vue opérationnel, de tels algorithmes sont évidemment indispensables à la mise en œuvre de nos modèles. Qui plus est, leur formulation nous fournira nombre d'ingrédients qui nous permettront d'obtenir une solution également efficace du problème d'apprentissage à la section 4. Les algorithmes que nous allons décrire ont fait l'objet de nombreux travaux et sont parfaitement documentés dans la littérature. Nous n'en ferons donc ici qu'un survol rapide, nous limitant à présenter les outils dont nous aurons besoin dans la suite.

3.1. Algorithmes uni-dimensionnels

Le problème de la reconnaissance dans des modèles doublement stochastiques dont la composante non observable est une chaîne de Markov a été étudié en [9], [13], [35], [42] et [48]. Notre présentation s'inspire de [18].

Considérons tout d'abord le problème de l'évaluation de la probabilité conjointe — ou vraisemblance — $\mathscr{L}_\tau(\psi_j) = \mathrm{P}(\omega^\tau = \psi_j,\ \bar{\mathrm{X}}_1^\tau)$ de l'état ψ_j au temps τ et de la séquence d'observations $\bar{\mathrm{X}}_1^\tau$. A la manière de Baum,

nous aborderons le problème en remarquant que nous disposons de la décomposition simple

$$(13)\quad \tilde{\mathscr{L}}_\tau(\psi_j) \doteq \mathrm{P}(\omega^\tau=\psi_j,\ \bar{\mathrm{X}}_1^{\mathrm{T}}) = \mathrm{P}(\omega^\tau=\psi_j,\ \bar{\mathrm{X}}_1^\tau)\,\mathrm{P}(\bar{\mathrm{X}}_{\tau+1}^{\mathrm{T}}/\omega^\tau=\psi_j) = \tilde{\mathscr{F}}_\tau(\psi_j)\,\tilde{\mathscr{B}}_\tau(\psi_j),$$

pour $j=1, \ldots, \vartheta$ et $1 \leqq \tau < \mathrm{T}$. Dans la littérature anglo-saxonne, les probabilités $\tilde{\mathscr{F}}_\tau(\psi_j)$ et $\tilde{\mathscr{B}}_\tau(\psi_j)$ sont généralement appelées probabilité *forward* et probabilité *backward* respectivement, pour des raisons qui vont apparaître dans un instant.

On vérifie aisément [18] que les probabilités forward peuvent être calculées par une récurrence *progressive*

$$(14)\quad \tilde{\mathscr{F}}_\tau(\psi_j) = \mathrm{P}_j p_j(\mathrm{X}^1) \qquad \tau=1,$$
$$= \textstyle\sum_i \tilde{\mathscr{F}}_{\tau-1}(\psi_i)\,\mathrm{P}_{ij}\,p_j(\mathrm{X}^\tau), \qquad \tau=2, \ldots, \mathrm{T},$$

tandis que les probabilités backward peuvent être calculées par une récurrence *régressive*

$$(15)\quad \tilde{\mathscr{B}}_\tau(\psi_j) = 1, \qquad \tau=\mathrm{T},$$
$$= \textstyle\sum_k \mathrm{P}_{jk}\,p_k(\mathrm{X}^{\tau+1})\,\tilde{\mathscr{B}}_{\tau+1}(\psi_k), \qquad \tau=\mathrm{T}-1, \ldots, 1.$$

Au vu de (13), l'algorithme forward-backward est tout à fait simple. Les probabilités $\tilde{\mathscr{F}}$ sont calculées par (14) et *mémorisées* au cours d'une passe progressive. Ensuite, au cours d'une passe régressive, les probabilités $\tilde{\mathscr{B}}$ sont calculées par (15) et le produit apparaissant dans le membre de droite de (13) peut être effectué, ce qui nous donne les probabilités attendues $\tilde{\mathscr{L}}_\tau(\psi_j)$ pour $j=1, \ldots, \vartheta$ et $\tau=\mathrm{T}, \ldots, 1$.

On peut observer que l'expression $\tilde{\mathscr{L}} = p(\mathrm{X}^1, \ldots, \mathrm{X}^{\mathrm{T}})$ que nous avons recontrée précédemment en (3) peut aussi s'écrire

$$\sum_j \tilde{\mathscr{L}}_\tau(\psi_j) = \sum_j \tilde{\mathscr{F}}_\tau(\psi_j)\,\tilde{\mathscr{B}}_\tau(\psi_j),$$

uniformément en τ. D'autre part, il est évident que (14) et (15) sont de complexité *linéaire* en T, c'est-à-dire que chaque exécution de la récurrence requiert un nombre fixé d'opérations qui ne dépend pas de τ. Il en découle que les deux récurrences (14) et (15) nous ont permis de nous affranchir de la complexité exponentielle qui caractérisait $\tilde{\mathscr{L}}$ en (3).

La représentation en graphe fournit à nouveau une interprétation intuitive évidente de l'algorithme forward-backward. Ainsi, par exemple, la probabilité $\tilde{\mathscr{F}}_{\tau_5}(\psi_4)$ est-elle la somme des probabilités d'émission de la sous-séquence $\mathrm{X}^{\tau_1}, \ldots, \mathrm{X}^{\tau_5}$ le long de tous les chemins distincts et menant au nœud $\omega^{\tau_5}=\psi_4$ lorsque le graphe est traversé de gauche à droite. Ceci est illustré à la figure 2. La probabilité backward, ainsi que la vraisemblance $\tilde{\mathscr{L}}_\tau(\psi_j)$ peuvent être interprétées de la même manière.

Levinson *et al.* [42] ont fait remarquer que la mise en œuvre des récurrences dans un programme d'ordinateur donnerait inévitablement lieu à des problèmes numériques de dépassement inférieur de capacité. Ceci s'explique par le fait que $\tilde{\mathscr{F}}_\tau(\psi_j) \overset{\tau}{\Rightarrow} 0$ et $\tilde{\mathscr{B}}_\tau(\psi_j) \overset{\mathrm{T}-\tau}{\Rightarrow} 0$ d'une manière exponentielle. Ces auteurs ont aussi suggéré un remède relativement heuristique. Plus récemment, Devijver [18] a montré qu'une reformulation du calcul en termes de probabilités *a posteriori* engendre un algorithme exempt de problèmes numériques. Compte tenu du fait que les deux formulations nous seront nécessaires à la section suivante, nous présentons brièvement la seconde solution.

Soit $\mathscr{L}_\tau(\psi_j) \doteq \mathrm{P}(\omega^\tau=\psi_j/\bar{\mathrm{X}}_1^\tau)$ la probabilité *a posteriori* de l'état $\omega^\tau=\psi_j$ conditionnellement à la séquence $\bar{\mathrm{X}}_1^{\mathrm{T}}$. Par analogie avec (13), le raisonnement poursuivi en [18] a pour point de départ la décomposition suivante : pour tout $1 \leqq \tau < \mathrm{T}$,

$$(16)\quad \mathscr{L}_\tau(\psi_j) = \mathrm{P}(\omega^\tau=\psi_j/\bar{\mathrm{X}}_1^{\mathrm{T}}) = \frac{\mathrm{P}(\omega^\tau=\psi_j,\ \bar{\mathrm{X}}_1^\tau)}{p(\bar{\mathrm{X}}_1^\tau)} \times \frac{p(\bar{\mathrm{X}}_{\tau+1}^{\mathrm{T}}/\omega^\tau=\psi_j)}{p(\bar{\mathrm{X}}_{\tau+1}^{\mathrm{T}}/\bar{\mathrm{X}}_1^\tau)} = \mathscr{F}_\tau(\psi_j)\,\mathscr{B}_\tau(\psi_j), \qquad j=1, \ldots, \vartheta.$$

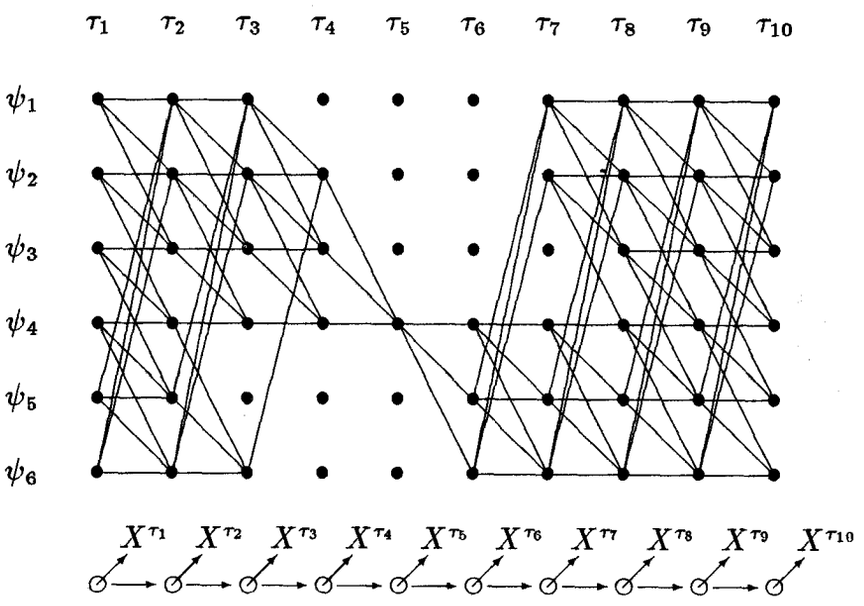


Fig. 2. – $\tilde{\mathscr{F}}_{\tau_5}(\psi_4)\,\tilde{\mathscr{B}}_{\tau_5}(\psi_4)$ **est la somme des probabilités d'émission de la sous-séquence** $\mathrm{X}^{\tau_1}, \ldots, \mathrm{X}^{\tau_{10}}$ **le long de tous les chemins distincts passant par le nœud** $\omega^{\tau_5}=\psi_4$.

Remarquons que $\mathscr{L}_\tau(\psi_j)$ et $\mathscr{F}_\tau(\psi_j)$ ont toutes deux la forme de probabilités *a posteriori* tandis que $\mathscr{B}_\tau(\psi_j)$ ne se prête à aucune interprétation naturelle ([5]).

La probabilité forward peut, à nouveau, être calculée au moyen d'une récurrence progressive simple,

$$(17)\quad \begin{aligned}\mathscr{F}_\tau(\psi_j) &= \mathscr{N}_\tau \mathrm{P}_j p_j(\mathrm{X}^1), && \tau=1\\ &= \mathscr{N}_\tau \textstyle\sum_i \mathscr{F}_{\tau-1}(\psi_i)\,\mathrm{P}_{ij} p_j(\mathrm{X}^\tau), && \tau=2,\ \ldots,\ \mathrm{T},\end{aligned}$$

où $\mathscr{N}_\tau$ joue le rôle de facteur de normalisation.

$$(18)\quad \begin{aligned}\mathscr{N}_\tau &= [\textstyle\sum_j \mathrm{P}_j p_j(\mathrm{X}^\tau)]^{-1}, && \tau=1\\ &= [\textstyle\sum_i \sum_j \mathscr{F}_{\tau-1}(\psi_i)\,\mathrm{P}_{ij} p_j(\mathrm{X}^\tau)]^{-1}, && \tau=2,\ \ldots,\ \mathrm{T}.\end{aligned}$$

La probabilité backward donne lieu à la décomposition

$$(19)\quad \mathscr{B}_\tau(\psi_j) = \sum_k \frac{\mathrm{P}_{jk} p_k(\mathrm{X}^{\tau+1})}{p(\mathrm{X}_{\tau+1}/\bar{\mathrm{X}}_1^\tau)} \times \frac{\mathrm{P}(\bar{\mathrm{X}}_{\tau+2}^{\mathrm{T}}/\omega^{\tau+1}=\psi_k)}{p(\bar{\mathrm{X}}_{\tau+2}^{\mathrm{T}}/\bar{\mathrm{X}}_1^{\tau+1})}$$

qui n'est autre que la récurrence régressive

$$(20)\quad \begin{aligned}\mathscr{B}_\tau(\psi_j) &= 1, && \tau=\mathrm{T},\\ &= \mathscr{N}_{\tau+1} \textstyle\sum_k \mathrm{P}_{jk} p_k(\mathrm{X}^{\tau+1})\,\mathscr{B}_{\tau+1}(\psi_k), && \tau=\mathrm{T}-1,\ \ldots,\ 1,\end{aligned}$$

où $\mathscr{N}_{\tau+1}$ est, à nouveau, le facteur de normalisation défini par (18).

Il est évident que les récurrences (17) et (20) exprimées en termes de probabilités *a posteriori* sont formellement identiques aux récurrences (14) et (15) exprimées en termes de vraisemblances, n'était la présence du facteur de normalisation $\mathscr{N}_\tau$. On remarquera aussi que la complexité du calcul — qui était de l'ordre $\mathscr{O}(\vartheta)$ en ce qui concerne (14)-(15) — semble être devenue $\mathscr{O}(\vartheta^2)$, de par la nécessité de calculer $\mathscr{N}_\tau$ en (18). Toutefois, une analyse plus fine montre que la récurrence progressive peut être mise en œuvre d'une manière plus efficace, à savoir, au moyen de l'algorithme suivant :

$$(21)\quad \begin{aligned}\mathscr{G}_\tau(\psi_j) &= \textstyle\sum_i \mathscr{F}_{\tau-1}(\psi_i)\,\mathrm{P}_{ij}, && \tau=2,\ \ldots,\ \mathrm{T},\\ \mathscr{F}'_\tau(\psi_j) &= \mathrm{P}_j p_j(\mathrm{X}^1), && \tau=1,\\ &= \mathscr{G}_\tau(\psi_j) p_j(\mathrm{X}^\tau), && \tau=2,\ \ldots,\ \mathrm{T},\\ \mathscr{N}_\tau &= [\textstyle\sum_j \mathscr{F}'_\tau(\psi_j)]^{-1}, && \tau=1,\ \ldots,\ \mathrm{T},\\ \mathscr{F}_\tau(\psi_j) &= \mathscr{N}_\tau \mathscr{F}'_\tau(\psi_j), && \tau=1,\ \ldots,\ \mathrm{T}.\end{aligned}$$

Dans cette formulation, $\mathscr{G}_\tau(\psi_j)$ est la probabilité *a priori* de $\omega^\tau=\psi_j$ *avant l'observation de* X^τ, soit $\mathrm{P}(\omega^\tau=\psi_j/\bar{\mathrm{X}}_1^{\tau-1})$. Présentée sous cette forme, la récurrence progressive n'est rien d'autre que la mise en œuvre la plus économique qui soit d'un algorithme proposé initialement, et indépendamment, par Abend [1] et Raviv [49] il y a de cela plus de 20 ans. (Devijver et Kittler [21] avaient déjà apporté de nombreuses simplifications à la méthode de Abend et Raviv, sans toutefois atteindre l'économie d'effort que traduit (21) qui est d'ordre $\mathscr{O}(\vartheta)$.)

A ce stade, un estimateur $\hat{\omega}^\tau$ de l'état du système au temps τ s'obtient par la règle du maximum de probabilité *a posteriori*,

$$\hat{\omega}^\tau=\psi_j \text{ si } \mathscr{L}_\tau(\psi_j)=\max_i \mathscr{L}_\tau(\psi_i) \qquad \tau=1,\ \ldots,\ \mathrm{T}.$$

Ceci n'est rien d'autre qu'une règle de décision de Bayes pour une fonction de coût (0,1). Les paramètres du modèle étant supposés connus, cette règle de décision est caractérisée par une probabilité d'erreur minimale [1, 49].

Avant de clore cette section, nous souhaitons encore mettre en évidence une relation qui sera utile dans la suite : en vertu de la définition de $\mathscr{L}_\tau(\psi_j)$ en (16), il vient

$$(22)\qquad \sum_j \mathscr{L}_\tau(\psi_j) = \sum_j \mathscr{F}_\tau(\psi_j)\,\mathscr{B}_\tau(\psi_j) = 1,$$

uniformément en τ.

3.2. Un algorithme bi-dimensionnel

Les méthodes de calcul développées ci-dessus constituent tout l'arsenal dont nous aurons besoin pour décrire la technique de reconnaissance proposée par Haslett [31] ainsi que nos algorithmes d'apprentissage de la section suivante. Supposons donc que la dépendance entre les états des pixels d'une image soit régie par un champ de Pickard, dans lequel, comme nous l'avons vu, les lignes et les colonnes sont des chaînes de Markov. Avec Haslett, nous émettons l'hypothèse simplificatrice suivante :

$$(23)\quad \begin{aligned}&\mathrm{P}(\lambda_{m,n}=\psi_j/\text{tous } \mathrm{X}_{k,l})\\ &\quad \approx \mathrm{P}(\lambda_{m,n}=\psi_j/\mathrm{X}_{k,l} : k=m \text{ ou } l=n)\\ &\quad \doteq {}_{\mathrm{H}}\mathscr{L}_{m,n}(\psi_j).\end{aligned}$$

En d'autres termes, ceci revient à négliger l'information relative à $\lambda_{m,n}$ que pourraient apporter les observations faites aux pixels qui ne se trouvent pas dans la ligne d'indice m ou la colonne d'indice n. C'est à ce stade que les propriétés d'indépendance conditionnelle de l'équation (13) trouvent leur utilité. En effet, elles nous permettent d'écrire

$$(24)\quad \begin{aligned}{}_{\mathrm{H}}\mathscr{L}_{m,n}(\psi_j) \propto\ & \mathrm{P}(\lambda_{m,n}=\psi_j/\bar{\mathrm{X}}_{m,1}^{m,n-1})\\ & \times \mathrm{P}(\lambda_{m,n}=\psi_j/\bar{\mathrm{X}}_{1,n}^{m-1,n})\\ & \times p(\mathrm{X}_{m,n}/\lambda_{m,n}=\psi_j)\\ & \times p(\bar{\mathrm{X}}_{m,n+1}^{m,\mathrm{N}}/\lambda_{m,n}=\psi_j)\\ & \times p(\bar{\mathrm{X}}_{m+1,n}^{\mathrm{M},n}/\lambda_{m,n}=\psi_j).\end{aligned}$$

On constate que les deux premiers facteurs du membre de droite de (24) ont la même forme que $\mathscr{G}_\tau(\psi_j)$ en (21), tandis que les deux derniers ont la même forme que $\mathscr{B}_\tau(\psi_j)$ en (20). Dès lors, en faisant à nouveau appel aux structures de chaîne de Markov le long des lignes et colonnes du champ de Pickard, ces quatre facteurs peuvent être calculés au moyen des récurrences uni-dimensionnelles de la section précédente. Dans le but de condenser l'écriture, nous posons ${}_m\mathscr{G}_n(\psi_j)$ pour $\mathscr{G}_n(\lambda_{m,n}=\psi_j)$ le long de la ligne

([5]) Le lecteur sera bien avisé de prendre note que les symboles caligraphiés, par exemple $\mathscr{L}$, $\mathscr{F}$, $\mathscr{B}$, accompagnés du signe *tilde* représentent des vraisemblances — ou probabilités conjointes — tandis que ceux qui en sont dépourvus représentent des probabilités *a posteriori*. Nous adhérons fidèlement à cette convention dans la suite.

d'indice m, ${}_n\mathscr{G}_m(\psi_j)$ pour $\mathscr{G}_m(\lambda_{m,n}=\psi_j)$ le long de la colonne d'indice n, et nous procédons de même avec les facteurs $\mathscr{B}$. De (24) nous tirons donc

$$(25)\quad {}_H\mathscr{L}_{m,n}(\psi_j) \propto {}_m\mathscr{G}_n(\psi_j)\,{}_n\mathscr{G}_m(\psi_j) \times p_j(X_{m,n})\,{}_m\mathscr{B}_n(\psi_j)\,{}_n\mathscr{B}_m(\psi_j).$$

L'équation (25) montre que la probabilité ${}_H\mathscr{L}_{m,n}(\psi_j)$ peut être calculée par deux applications orthogonales de l'algorithme forward-backward, une fois le long de la ligne d'indice m et une fois le long de la colonne d'indice n. Cela étant fait, l'estimateur de $\lambda_{m,n}$ qui minimize la probabilité d'erreur est obtenu par $\text{argmax}^j\{{}_H\mathscr{L}_{m,n}(\psi_j)\}$ comme à la section précédente. Le lecteur trouvera en [31], et à la section 5 ci-après, des résultats expérimentaux illustrant les performances que l'on peut attendre de cette méthode.

La technique que nous venons de décrire n'est qu'une des nombreuses méthodes possibles applicables au modèle d'image doublement stochastique dont la composante markovienne est un champ de Pickard. D'autres techniques [15, 17, 20 et 39] applicables à la classe plus large des réseaux de Markov peuvent, en principe, être implantées en temps réel grâce à leur exploitation astucieuse de la causalité de (8). Leur mise en œuvre requiert toutefois l'exécution d'un nombre considérable d'opérations par période d'échantillonnage. En ce qui concerne la classe la plus générale, à savoir celle des champs de Markov, on se trouve devant un éventail de méthodes, nécessairement itératives, parmi lesquelles l'*iterated conditional mode* de Besag [8], et l'algorithme du *recuit simulé* proposé initialement par les frères Geman [26] (*voir* aussi [12], [28] et [54]) occupent les extrêmes de l'échelle de complexité. En tout état de cause, notre intérêt pour la méthode décrite ci-dessus procède moins de la relative simplicité de sa mise en œuvre que de la possibilité qu'elle va nous offrir de mettre au point des algorithmes d'apprentissage.

4. Algorithmes d'apprentissage

4.1. Considérations préliminaires

Dans la section qui précède, nous avons passé en revue des algorithmes permettant d'exploiter de manière efficace l'information condensée dans un modèle markovien d'image, en supposant connus les paramètres du modèle. Nous verrons sous peu, qu'en dépit des apparences, il ne s'agissait pas d'un détour sur le chemin qui va nous conduire à des algorithmes d'apprentissage.

Afin d'éviter toute confusion possible, il nous paraît indispensable de circonscrire de manière précise, le type d'apprentissage qui va nous occuper. Dans certains contextes, l'apprentissage se réduit à un problème classique d'estimation statistique. Ainsi en va-t-il, par exemple, de la lecture optique. Dans ce domaine, le nombre d'états est fixé *a priori* (57 pour un alphabet composé de 26 majuscules, 26 minuscules, 4 signes d'accentuation et ponctuation et un « blanc »), les probabilités initiales et de transition de la chaîne de Markov modélisant le langage sont estimées par les fréquences correspondantes dans le langage considéré. Quant aux distributions conditionnelles, elles peuvent être estimées, pour chaque caractère, sur la base d'un échantillon de signaux correspondant au caractère en question. Dans la terminologie de la reconnaissance des formes, un tel « apprentissage » est dit *supervisé*. Cette technique à été largement exploitée dans le passé [44, 49, 52, 53], mais il faut convenir que les résultats obtenus ne répondirent pas toujours aux espoirs des expérimentateurs [29, 51].

Comme il est toujours utile de tirer profit des leçons du passé, nous observons que l'apprentissage supervisé confie à l'expérimentateur le soin de spécifier *a priori* tous les détails du modèle et repose sur l'espoir, toujours aléatoire, que *les données s'accorderont bien au modèle proposé*. Ainsi peut-on trouver sous la plume de David Forney, en 1973, le but déclaré d'illustrer « *different sorts of problems that can be made to fit such a model well* » [25].

Notre démarche sera diamétralement opposée à celle de l'apprentissage supervisé. En effet, nous définirons une famille de modèles par un minimum de spécifications *a priori* et nous rechercherons parmi l'ensemble des modèles, celui qui *s'accorde le mieux aux données expérimentales*. Par exemple, si nous ne pourrons éviter de pré-spécifier le nombre et la nature des états, nous nous garderons de définir *a priori* ce que doit être chacun des états possibles. Par voie de conséquence, notre technique d'apprentissage sera de type *non supervisé*.

Comme pour la partie « reconnaissance » de la section 3, la première méthode que nous appliquerons aux modèles d'image qui se fondent sur les champs de Pickard s'inspire fortement des techniques utilisées en reconnaissance de la parole. A nouveau, un survol rapide de ces techniques s'avère-t-il donc indispensable.

4.2. Apprentissage de modèles uni-dimensionnels

Considérons à nouveau un modèle doublement stochastique dont la composante non observable est une chaîne de Markov. Remarquons que, à ce stade, nous n'avons spécifié en aucune manière les distributions conditionnelles $p_j(X)$, $j=1, \ldots, \vartheta$. Dans le but de simplifier la présentation, nous supposerons momentanément que la variable aléatoire X est discrète et prend ses valeurs dans l'ensemble $\Xi=\{\xi_1, \ldots, \xi_K\}$, avec la probabilité $p_j(\xi_k)=P(X^\tau=\xi_k/\omega^\tau=\psi_j)$, $1\leq j\leq\vartheta$, $1\leq k\leq K$, $1\leq\tau\leq T$. En tout état de cause, et sans entrer dans des considérations théoriques par ailleurs fort importantes, nous supposerons constamment que ces distributions sont *identifiables* [55, 56].

Nous avons établi à la section 2.1 que la vraisemblance $\tilde{\mathscr{L}}$ d'une séquence $\bar{X}_1^T=\{X^1, \ldots, X^T\}$ d'observations est donnée par (3), qui, combinée avec (13) et (15), devient

$$(26)\quad \tilde{\mathscr{L}}=p(X^1, \ldots, X^T) = \sum_{i=1}^{\vartheta} \tilde{\mathscr{F}}_\tau(\psi_i)\,\tilde{\mathscr{B}}_\tau(\psi_i) = \sum_{i=1}^{\vartheta}\sum_{j=1}^{\vartheta} \tilde{\mathscr{F}}_\tau(\psi_i)\,P_{ij}\,p_j(X^{\tau+1})\,\tilde{\mathscr{B}}_{\tau+1}(\psi_j),$$

uniformément en $\tau > 1$. En supposant donnés le nombre d'états ϑ et la séquence d'observations $\bar{\mathbf{X}}_1^T$, nous allons considérer $\tilde{\mathscr{L}}$ comme une fonction des paramètres $\{P_i, P_{ij}, p_j(\xi_k)\}$, et formuler notre premier problème d'apprentissage comme celui de l'estimation des paramètres qui maximisent la vraisemblance $\tilde{\mathscr{L}}$ de $\bar{\mathbf{X}}_1^T$, sous les contraintes $\sum_i P_i = 1$, $\sum_j P_{ij} = 1$, $\forall i$, et $\Sigma_k p_j(\xi_k) = 1$, $\forall j$. (Nous verrons dès la section suivante qu'il n'y a pas lieu de s'inquiéter des problèmes numériques liés à l'évaluation d'une vraisemblance.)

Notre problème d'optimisation se prête bien à l'utilisation de la technique des multiplicateurs de Lagrange. Nous formons dès lors la fonction auxiliaire

$$\Pi = \tilde{\mathscr{L}} + \alpha\Big[\sum_i P_i - 1\Big] + \sum_i \beta_i \Big[\sum_j P_{ij} - 1\Big] + \sum_l \gamma_l \Big[\sum_k p_l(\xi_k) - 1\Big], \tag{27}$$

où α, β_i et γ_l, $i, l = 1, \ldots, \vartheta$ sont les multiplicateurs de Lagrange. L'équation (27) montre que l'optimisation par rapport à une famille de paramètres peut être effectuée indépendamment de l'optimisation par rapport aux autres familles de paramètres. En conséquence, nous nous bornerons ici à traiter le problème de l'optimisation par rapport aux probabilités de transition P_{ij}, les autres cas pouvant être traités de manière parfaitement analogue.

En égalant à zéro la dérivée partielle de Π par rapport à P_{ij} nous obtenons

$$\frac{\partial \Pi}{\partial P_{ij}} = \frac{\partial \tilde{\mathscr{L}}}{\partial P_{ij}} + \beta_i = 0. \tag{28}$$

Une multiplication par P_{ij} fournit

$$P_{ij}\, \partial \tilde{\mathscr{L}} / \partial P_{ij} + \beta_i P_{ij} = 0, \tag{29}$$

soit, en vertu de la contrainte $\sum_j P_{ij} = 1$,

$$\beta_i = -\sum_j P_{ij}\, \partial \tilde{\mathscr{L}} / \partial P_{ij}. \tag{30}$$

En substituant β_i de (30) en (29) nous obtenons la solution sous la forme d'une équation implicite

$$P_{ij} = \frac{P_{ij}\, \partial \tilde{\mathscr{L}} / \partial P_{ij}}{\sum_j P_{ij}\, \partial \tilde{\mathscr{L}} / \partial P_{ij}}. \tag{31}$$

En ce qui concerne les autres paramètres, on obtient de manière similaire

$$P_i = \frac{P_i\, \partial \tilde{\mathscr{L}} / \partial P_i}{\sum_j P_j\, \partial \tilde{\mathscr{L}} / \partial P_j} \tag{32}$$

et

$$p_j(\xi_k) = \frac{p_j(\xi_k)\, \partial \tilde{\mathscr{L}} / \partial p_j(\xi_k)}{\sum_k p_j(\xi_k)\, \partial \tilde{\mathscr{L}} / \partial p_j(\xi_k)}. \tag{33}$$

Examinons à présent le problème de l'évaluation de la dérivée partielle $\partial \tilde{\mathscr{L}} / \partial P_{ij}$. En premier lieu, nous exprimons $\tilde{\mathscr{L}}$ comme la somme des vraisemblances de $\bar{\mathbf{X}}_1^T$ le long de tous les chemins possibles distincts dans la représentation en graphe du modèle doublement stochastique.

$$\tilde{\mathscr{L}} = \sum_{\tau=1}^{T-1} \sum_{k=1} \sum_{l=1} P(\bar{\mathbf{X}}_1^T, \omega^\tau = \psi_k, \omega^{\tau+1} = \psi_l) \tag{34}$$

Nous remplaçons chaque terme de la triple somme par sa décomposition forward-backward (13) dans laquelle la récurrence régressive est substituée à la probabilité backward. Ce faisant, nous obtenons

$$\tilde{\mathscr{L}} = \sum_{\tau=1}^{T-1} \sum_{k=1} \sum_{l=1} \mathscr{F}_\tau(\psi_k) P_{kl} p_l(X^{\tau+1}) \mathscr{B}_{\tau+1}(\psi_l). \tag{35}$$

Il est clair que le terme $\mathscr{F}_\tau(\psi_k) P_{kl} p_l(X^{\tau+1}) \mathscr{B}_{\tau+1}(\psi_l)$ en (35) apportera à $\partial \tilde{\mathscr{L}} / \partial P_{ij}$ une contribution égale à $\mathscr{F}_\tau(\psi_k) p_l(X^{\tau+1}) \mathscr{B}_{\tau+1}(\psi_l)$ si, et seulement si, $k = i$ et $l = j$. Il s'ensuit que

$$\frac{\partial \tilde{\mathscr{L}}}{\partial P_{ij}} = \sum_{\tau=1}^{T-1} \mathscr{F}_\tau(\psi_i) p_j(X^{\tau+1}) \mathscr{B}_{\tau+1}(\psi_j),$$

et

$$P_{ij} \frac{\partial \tilde{\mathscr{L}}}{\partial P_{ij}} = \sum_{\tau=1}^{T-1} \mathscr{F}_\tau(\psi_i) P_{ij} p_j(X^{\tau+1}) \mathscr{B}_{\tau+1}(\psi_j). \tag{36}$$

En substituant ces valeurs en (31) et en procédant de la même manière avec les équations (32) et (33), nous obtenons les *formules de ré-estimation* [42],

$$P_i = \frac{\tilde{\mathscr{F}}_1(\psi_i) \tilde{\mathscr{B}}_1(\psi_i)}{\sum_j \tilde{\mathscr{F}}_1(\psi_j) \tilde{\mathscr{B}}_1(\psi_j)} \tag{37}$$

$$= \frac{P(X^1, \ldots, X^T, \omega^1 = \psi_i)}{\sum_j P(X^1, \ldots, X^T, \omega^1 = \psi_j)}, \tag{38}$$

$$P_{ij} = \frac{\sum_\tau \tilde{\mathscr{F}}_\tau(\psi_i) P_{ij} p_j(X^{\tau+1}) \tilde{\mathscr{B}}_{\tau+1}(\psi_j)}{\sum_j \sum_\tau \tilde{\mathscr{F}}_\tau(\psi_i) P_{ij} p_j(X^{\tau+1}) \tilde{\mathscr{B}}_{\tau+1}(\psi_j)} \tag{39}$$

$$= \frac{\sum_\tau \tilde{\mathscr{F}}_\tau(\psi_i) P_{ij} p_j(X^{\tau+1}) \tilde{\mathscr{B}}_{\tau+1}(\psi_j)}{\sum_\tau \tilde{\mathscr{F}}_\tau(\psi_i) \tilde{\mathscr{B}}_\tau(\psi_i)} \tag{40}$$

$$= \frac{\sum_\tau P(X^1, \ldots, X^T, \omega^\tau = \psi_i, \omega^{\tau+1} = \psi_j)}{\sum_\tau P(X^1, \ldots, X^T, \omega^\tau = \psi_i)}, \tag{41}$$

$$p_j(\xi_k) = \frac{\sum_{\tau \mid X^\tau = \xi_k} \tilde{\mathscr{F}}_\tau(\psi_j) \tilde{\mathscr{B}}_\tau(\psi_j)}{\sum_\tau \tilde{\mathscr{F}}_\tau(\psi_j) \tilde{\mathscr{B}}_\tau(\psi_j)} \tag{42}$$

$$= \frac{\sum_{\tau \mid X^\tau = \xi_k} P(X^1, \ldots, X^\tau, \ldots, X^T, \omega^\tau = \psi_j)}{\sum_\tau P(X^1, \ldots, X^T, \omega^\tau = \psi_j)}. \tag{43}$$

Les équations (37)-(43) nous fournissent le principe de la méthode itérative de ré-estimation des paramètres P_i, P_{ij} et $p_j(\xi_k)$.

Du point de vue opérationnel, et en supposant que l'on dispose au préalable de valeurs initiales — *voir* à ce sujet la section 5 —, une itération requiert le calcul des probabilités $\tilde{\mathscr{F}}$ et $\tilde{\mathscr{B}}$ obtenues avec les valeurs courantes des paramètres, et la mise à jour, ou ré-estimation de ces paramètres au moyen des équations (37), (39) et (42). Cette technique est itérée jusqu'à la convergence en un extremum de la fonction $\tilde{\mathscr{L}}$.

Les équations (38), (40) et (43) montrent que nous avons bien affaire à un algorithme de type EM [50] et nous fournissent une interprétation simple des

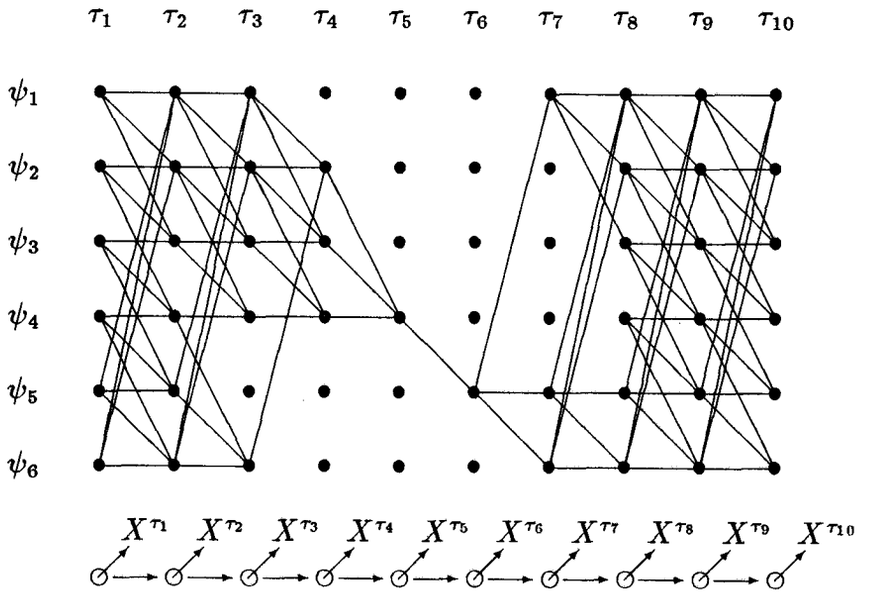


Fig. 3. – Chaque terme du numérateur de la formule de ré-estimation de P_{ij} est la probabilité d'observer la séquence $X^{\tau}{}_1, \ldots, X^{\tau}{}_{10}$ et de faire la transition spécifiée.

expressions mathématiques intervenant dans les formules de ré-estimation. En effet, dans le cas de l'équation (40) par exemple, nous voyons que le numérateur et le dénominateur sont les sommes des probabilités de la séquence $X^1, \ldots, X^T$ et de la transition $\psi_i - \psi_j$ d'une part, et de toute transition issue de ψ_i d'autre part. Un des termes de la somme du numérateur de (40) est illustré à la figure 3 pour le modèle introduit à la section 2.1 et $(i, j) = (4,5)$.

L'analogie avec l'algorithme EM classique – pour des observations indépendantes et identiquement distribuées, en d'autres mots, compte non tenu de la dépendance Markovienne (voir à ce propos la section 6.4 de [24]) – est encore plus apparente lorsque l'on suppose que la variable X est continue et distribuée selon ϑ lois normales. Dans ce cas, ces lois sont spécifiées par leurs paramètres, notés θ_j, $j = 1, \ldots, \vartheta$, « espérance mathématique » et « matrice de covariance » pour lesquels des formules de ré-estimation peuvent être établies par la méthode que nous venons d'examiner. La formule (37) n'est pas affectée tandis que (39) et (42) deviennent

$$(44) \quad P_{ij} = \frac{\sum_\tau \tilde{\mathcal{F}}_\tau(\psi_i)\, P_{ij}\, p(X^{\tau+1} / \psi_j, \hat{\theta}_j)\, \tilde{\mathcal{B}}_{\tau+1}(\psi_j)}{\sum_j \sum_\tau \tilde{\mathcal{F}}_\tau(\psi_i)\, P_{ij}\, p(X^{\tau+1} / \psi_j, \hat{\theta}_j)\, \tilde{\mathcal{B}}_{\tau+1}(\psi_j)}$$

où $p(X^{\tau+1} / \psi_j, \hat{\theta}_j) \sim N(X^{\tau+1}; \hat{\mu}_j, \hat{\Sigma}_j)$ avec

$$(45) \quad \hat{\mu}_j = \frac{\sum_\tau \tilde{\mathcal{F}}_\tau(\psi_j)\, \tilde{\mathcal{B}}_\tau(\psi_j)\, X^\tau}{\sum_\tau \tilde{\mathcal{F}}_\tau(\psi_j)\, \tilde{\mathcal{B}}_\tau(\psi_j)} = \frac{\sum_\tau P(\omega^\tau = \psi_j, X^1, \ldots, X^T)\, X^\tau}{\sum_\tau P(\omega^\tau = \psi_j, X^1, \ldots, X^T)}$$

et

$$(46) \quad \hat{\Sigma}_j = \frac{\sum_\tau \tilde{\mathcal{F}}_\tau(\psi_j)\, \tilde{\mathcal{B}}_\tau(\psi_j)\, (X^\tau - \hat{\mu}_j)\, (X^\tau - \hat{\mu}_j)'}{\sum_\tau \tilde{\mathcal{F}}(\psi_j)\, \tilde{\mathcal{B}}_\tau(\psi_j)}$$

où $(X^\tau - \hat{\mu}_j)'$ est le transposé du vecteur $(X^\tau - \hat{\mu}_j)$.

La convergence de l'algorithme EM pour des variables aléatoires indépendantes et identiquement distribuées a été étudiée par trop d'auteurs pour les énumérer ici. Le lecteur intéressé consultera utilement l'article de Redner et Walker qui contient 162 références [50]. Rappelons toutefois que la convergence vers un optimum global n'est pas garantie : un optimum local peut être atteint, aussi bien qu'une solution singulière à la limite de l'espace des paramètres admissibles.

En ce qui concerne l'algorithme que nous venons d'évoquer, mentionnons que Baum *et al.* ont établi les conditions de convergence pour divers types de distributions conditionnelles uni-variées discrètes et continues : distribution de Poisson sur les entiers non négatifs, distribution binomiale sur les entiers, loi gamma et loi normale [5]. Liporace a généralisé ces résultats à des distributions multivariées normales et de Cauchy [43]. Les expérimentations de Levinson *et al.* [42] ont montré que toute asymétrie de la matrice des probabilités de transition influence défavorablement la vitesse de convergence et que le choix des valeurs initiales est tout à fait critique. Différents aspects des questions liées à la convergence seront illustrés à la section 5.

4.3. Apprentissages de modèles bi-dimensionnels

4.3.1. *L'algorithme EM*

A la section précédente, nous avons vu comment estimer les paramètres du modèle uni-dimensionnel sur la base de l'observation d'une séquence $\bar{X}_1^T$. Dans la plupart des applications, on souhaiterait évidemment fonder l'estimation sur plusieurs séquences, en réalité, sur le plus grand nombre possible de séquences, de manière à améliorer la précision des estimateurs. Nous allons développer cette idée dans le contexte du modèle d'image où la dépendance entre états est modélisée par un champ de Pickard.

Notre objectif est d'estimer les paramètres nécessaires à l'utilisation de (25) en reconnaissance. En d'autres mots, nous aurons à estimer les paramètres de la chaîne de Markov qui engendre lignes et colonnes du

champ de Pickard. En principe, une simple généralisation de l'algorithme de la section précédente devrait faire notre affaire, pourvu que nous puissions garantir de surcroît la *réversibilité* de la chaîne en question. En vertu de la remarque faite à la section 2.1, nous remettrons à plus tard cette difficulté supplémentaire en supposant, dans un premier temps, avoir affaire à un champ *binaire*.

Dans le but de simplifier l'écriture de ce qui suit, nous écrivons Y_m pour la réalisation de la ligne d'indice m, $1 \leqq m \leqq M$, et Z_n pour la réalisation de la colonne d'indice n, $1 \leqq n \leqq N$. Dans la notation de la section 3.2, $Y_m \equiv (\bar{X}_{m,1}^{m,N})$, et $Z_n \equiv (\bar{X}_{1,n}^{M,n})$. Dans ce qui suit, nous allons prendre quelque liberté supplémentaire par rapport au modèle théorique en émettant l'hypothèse simplificatrice que lignes et colonnes d'observations sont mutuellement indépendantes, ce que, bien évidemment, elles ne sont pas. Il s'ensuit que la *pseudo-vraisemblance* [7] de l'image entière est donnée par $[\prod_{m,n} P(Y_m) P(Z_n)]^{1/2}$. Comme nous nous proposons de déterminer les paramètres qui maximisent cette pseudo-vraisemblance, nous pouvons tout aussi bien prendre le carré de cette fonction objectif comme critère d'optimisation. De plus, nous allons voir sous peu que les lignes Y et colonnes Z de l'image sont traitées exactement de la même manière (6). D'où, pour simplifier au maximum la présentation, nous allons examiner le problème qui consiste à maximiser la fonction objectif

$$\mathbf{\Pi} = \prod_{m=1}^{M} P(Y_m)$$

par rapport aux probabilités de transition P_{ij}, sous la contrainte $\sum_j P_{ij} = 1$. Le lecteur n'aura aucune peine à vérifier que, comme à la section 4.2, chaque série de paramètres – (i) probabilités initiales, (ii) probabilités de transition et (iii) distributions conditionnelles, – donne lieu à un problème d'optimisation qui peut être traité de manière indépendante.

Soit donc à déterminer un extrémum de la fonction auxiliaire

$$\mathbf{\Pi} + \rho_i (\sum_j P_{ij} - 1),$$

où ρ_i est un multiplicateur de Lagrange, $i = 1, \ldots, 9$. En égalant les dérivées partielles à zéro et en éliminant les multiplicateurs nous obtenons

$$P_{ij} = \frac{P_{ij} (\partial \mathbf{\Pi} / \partial P_{ij})}{\sum_j P_{ij} (\partial \mathbf{\Pi} / \partial P_{ij})}. \tag{47}$$

Comme $\mathbf{\Pi}$ est un produit de facteurs, nous pouvons écrire

$$\frac{\partial \mathbf{\Pi}}{\partial P_{ij}} = \sum_{m=1}^{M} \prod_{k \neq m} P(Y_k) \frac{\partial P(Y_m)}{\partial P_{ij}} = \mathbf{\Pi} \sum_{m=1}^{M} \frac{\partial P(Y_m)}{\partial P_{ij}} \frac{1}{P(Y_m)} \tag{48}$$

(6) En fait, pour des images carrées, (M = N), la mise en œuvre la plus simple de la technique que nous allons décrire consiste à ajouter au bas de l'image initiale une copie de celle-ci obtenue après rotation de 90° et à traiter, le long des lignes uniquement, l'image 2M × M ainsi obtenue.

Par substitution de (48) en (47) le facteur $\mathbf{\Pi}$, commun au numérateur et au dénominateur, peut être simplifié et il nous reste

$$P_{ij} = \frac{\sum_{m=1}^{M} P_{ij} (\partial P(Y_m) / \partial P_{ij}) (1 / P(Y_m))}{\sum_{m=1}^{M} \sum_j P_{ij} (\partial P(Y_m) / \partial P_{ij}) (1 / P(Y_m))} \tag{49}$$

A ce stade, nous pouvons faire usage des résultats de la section précédente où nous avons établi que

$$P_{ij} \frac{\partial P(Y_m)}{\partial P_{ij}} = \sum_{n=1}^{N-1} {}_m\mathscr{F}_n(\psi_i) P_{ij} p_j(X_{m,n+1}) {}_m\mathscr{B}_{n+1}(\psi_j) = \sum_{n=1}^{N-1} P(Y_m, \lambda_{m,n} = \psi_i, \lambda_{m,n+1} = \psi_j), \tag{50}$$

où ${}_m\mathscr{F}_n$ et ${}_m\mathscr{B}_n$ doivent être interprétés comme à la section 3.2. De (50) nous tirons

$$\sum_j P_{ij} \frac{\partial P(Y_m)}{\partial P_{ij}} = \sum_{n=1}^{N-1} P(Y_m, \lambda_{m,n} = \psi_i). \tag{51}$$

Après avoir procédé aux substitutions de (50) et (51) dans (49), nous remarquons que les divisions de chaque terme par $P(Y_m)$ ont pour effet de transformer les probabilités conjointes dans les membres de droite de (50) et (51) en probabilités *a postériori*. C'est de cette manière que *nous allons disposer* à nouveau *d'un algorithme exempt de problème numérique de dépassement inférieur de capacité*. Finalement, nous obtenons la formule de ré-estimation des probabilités de transition

$$P_{ij} = \frac{\left\{ \sum_m \sum_n {}_m\mathscr{F}_n(\psi_i) {}_m\mathscr{N}_{n+1} P_{ij} p_j(X_{m,n+1}) {}_m\mathscr{B}_{n+1}(\psi_j) \right\}}{\left\{ \sum_m \sum_n \sum_j {}_m\mathscr{F}_n(\psi_i) {}_m\mathscr{N}_{n+1} P_{ij} p_j(X_{m,n+1}) {}_m\mathscr{B}_{n+1}(\psi_j) \right\}} \tag{52}$$

$$= \frac{\left\{ \sum_m \sum_n {}_m\mathscr{F}_n(\psi_i) {}_m\mathscr{N}_{n+1} P_{ij} p_j(X_{m,n+1}) {}_m\mathscr{B}_{n+1}(\psi_j) \right\}}{\sum_m \sum_n {}_m\mathscr{F}_n(\psi_i) {}_m\mathscr{B}_n(\psi_i)} \tag{53}$$

où les probabilités $\mathscr{F}$ et $\mathscr{B}$ (et le facteur de normalisation $\mathscr{N}$) sont calculées par les récurrences (17)-(18) [ou, de préférence (21)] et (20) respectivement.

La comparaison des équations (52)-(53) et (39)-(40) est instructive. Elle montre bien que, au problème de stabilité numérique près, l'estimateur bi-dimensionnel s'obtient par une simple méthode d'accumulation le long des lignes (et des colonnes) des probabilités intervenant au numérateur et au dénominateur de (39). Cette méthode d'accumulation trouvera une justification tout à fait intuitive à la section suivante.

Il va pratiquement sans dire que les formules de ré-estimation des autres paramètres du modèle sont établies selon le même principe d'accumulation. Par souci de complétude nous les donnons ci-dessous sous l'hypothèse que $p_i(X) \sim N(X;\ \mu_i,\ \Sigma_i)$.

$$(54) \qquad P_i = \frac{\sum_m {}_m\mathcal{F}_1(\psi_i)\,{}_m\mathcal{B}_1(\psi_i)}{\sum_m \sum_i {}_m\mathcal{F}_1(\psi_i)\,{}_m\mathcal{B}_1(\psi_i)},$$

$$(55) \qquad \hat{\mu}_i = \frac{\sum_m \sum_n {}_m\mathcal{F}_n(\psi_i)\,{}_m\mathcal{B}_n(\psi_i)\,X_{m,n}}{\sum_m \sum_n {}_m\mathcal{F}_n(\psi_i)\,{}_m\mathcal{B}_n(\psi_i)},$$

$$(56) \qquad \hat{\Sigma}_i = \frac{\sum_m \sum_n {}_m\mathcal{F}_n(\psi_i)\,{}_m\mathcal{B}_n(\psi_i)\,(X_{m,n}-\hat{\mu}_i)(X_{m,n}-\hat{\mu}_i)'}{\sum_m \sum_n {}_m\mathcal{F}_n(\psi_i)\,{}_m\mathcal{B}_n(\psi_i)}.$$

Ces formules de ré-estimation sont mises en œuvre de la même manière que (37)-(42) à la section 4.2 : les probabilités $\mathcal{F}$ et $\mathcal{B}$ sont évaluées par l'algorithme forward-backward en faisant usage des valeurs courantes des paramètres. Elles sont ensuite utilisées dans (52)-(56) pour calculer les mises à jour de ces paramètres. Le processus est répété jusqu'à convergence en un maximum (local) de la fonction de pseudo-vraisemblance.

Il y a lieu de garder présent à l'esprit le fait que l'usage de (52)-(56) devrait, en principe, être limité aux modèles d'images dans lesquels le champ de Pickard est supposé binaire. En effet, nous n'avons pas tenu compte jusqu'ici de la contrainte de réversibilité des chaînes de Markov engendrant les lignes et colonnes du champ de Pickard. Il nous était loisible d'agir de la sorte en vertu de la propriété de réversibilité des chaînes binaires.

Nous n'avons pas connaissance de contraintes simples qui entraînent la réversibilité d'une chaîne définie sur un espace d'états multi-valué, si ce n'est celle de la symétrie de la matrice des probabilités de transition. Eussions-nous imposé la contrainte supplémentaire $P_{ij} = P_{ji},\ \forall\, i \neq j$, dans notre problème d'optimisation, la formule de ré-estimation de P_{ij} eût été

$$(57) \qquad P_{ii} = \frac{\sum_m \sum_n {}_m\mathcal{F}_n(\psi_i)\,{}_m\mathcal{N}_{n+1}\,P_{ii}\,p_i(X_{m,n+1})\,{}_m\mathcal{B}_{n+1}(\psi_i)}{\sum_m \sum_n \left\{ {}_m\mathcal{F}_n(\psi_i)\,{}_m\mathcal{B}_n(\psi_i) - \sum_{j\neq i} (P_{ij}/P_{jj})\,{}_m\mathcal{F}_n(\psi_j)\,{}_m\mathcal{N}_{n+1}\,P_{jj}\,p_j(X_{m,n+1})\,{}_m\mathcal{B}_{n+1}(\psi_j) \right\}},$$

pour les éléments diagonaux de la matrice, et

$$(58) \qquad P_{ij} = \frac{\sum_m \sum_n \left\{ {}_m\mathcal{F}_n(\psi_i)\,{}_m\mathcal{N}_{n+1}\,P_{ij}\,p_j(X_{m,n+1})\,{}_m\mathcal{B}_{n+1}(\psi_j) - (P_{ij}/P_{jj})\,{}_m\mathcal{F}_n(\psi_j)\,{}_m\mathcal{N}_{n+1}\,P_{jj}\,p_j(X_{m,n+1})\,{}_m\mathcal{B}_{n+1}(\psi_j) \right\}}{\sum_m \sum_n \left\{ {}_m\mathcal{F}_n(\psi_i)\,{}_m\mathcal{B}_n(\psi_i) - \sum_{j\neq i} (P_{ij}/P_{jj})\,{}_m\mathcal{F}_n(\psi_j)\,{}_m\mathcal{N}_{n+1}\,P_{jj}\,p_j(X_{m,n+1})\,{}_m\mathcal{B}_{n+1}(\psi_j) \right\}}$$

pour les éléments non diagonaux. [Il y a lieu de lire (57) et (58) en adoptant la convention $0/0 = 0$.] Dans l'algorithme d'apprentissage pour un champ de Pickard multi-valué, il y a lieu de substituer (57) et (58) à (52). Comme la contrainte de symétrie de la matrice de transition n'affecte pas les autres paramètres du modèle, on peut vérifier facilement que les équations (54)-(56) restent inchangées.

4.3.2. *L'algorithme DD*

L'algorithme EM que nous venons de décrire est d'un intérêt théorique indéniable. Cependant sa mise en œuvre peut poser des problèmes de temps calcul lorsqu'on a affaire à des images de grande taille et que l'on ne dispose pas de « bonnes » valeurs initiales pour les paramètres à estimer. Il est possible de remédier partiellement à cet état de choses en ayant recours à un algorithme de type DD (Decision Directed), c'est-à-dire une version *approximative* et simplifiée du précédent dont on peut légitimement attendre qu'elle converge plus rapidement [24].

Il existe une multitude de variations sur le thème « DD ». Aussi, importe-t-il de préciser celle que nous avons adoptée : nous présentons ci-dessous l'approximation DD de l'algorithme défini par les équations (53)-(56).

Considérons d'abord l'équation (54) qui, étendue aux lignes et colonnes de l'image, peut s'écrire ([7])

$$(59) \qquad P_i = \frac{\sum_{m=1}^{M} P(\lambda_{m,1} = \psi_i / Y_m) + \sum_{n=1}^{N} P(\lambda_{1,n} = \psi_i / Z_n)}{M+N},$$

où la forme du dénominateur résulte de la substitution de (22) en (54). Si nous introduisons dans (59) l'approximation

$$(60) \qquad P(\lambda_{m,1} = \psi_i / Y_m) \approx \begin{cases} 1, & \text{si } P(\lambda_{m,1} = \psi_i / Y_m) = \max_j P(\lambda_{m,1} = \psi_j / Y_m) \\ 0, & \text{sinon} \end{cases}$$

et procédons de même avec $P(\lambda_{1,n} = \psi_i / Z_n)$, nous constatons que l'estimateur de P_i en (59) n'est autre chose que la *fréquence empirique* avec laquelle l'état ψ_i est assigné aux pixels de la première ligne et de la première colonne de l'image. Nous exprimons ceci

([7]) Au risque de nous répéter, il nous paraît indispensable d'insister sur le fait que le membre de droite d'une formule de ré-estimation telle que (59) est évalué au moyen de l'algorithme forward-backward utilisant les valeurs courantes des paramètres, tandis que le membre de gauche fournit la valeur mise à jour à utiliser à l'itération d'apprentissage suivante.

en faisant usage de la *fonction indicatrice* $I_{[(\,.\,)]}$ de l'évènement (.) :

$$P_i = \frac{1}{M+N-1} \times \left(\sum_{n=1}^{N} I_{[\lambda_{1,n}=\psi_i]} + \sum_{m=2}^{M} I_{[\lambda_{m,1}=\psi_i]} \right), \quad i=1,\ldots,\vartheta. \tag{61}$$

Il est évident que le même raisonnement peut s'appliquer aux équations (55) et (56) qui font intervenir les mêmes probabilités conditionnelles pour les états. En supposant que l'observation $X_{m,n}$ est uni-variée et distribuée selon ϑ lois normales de moyennes μ_i et d'écarts types σ_i, $i=1,\ldots,\vartheta$, nous obtenons les formules de ré-estimation suivantes

$$\mu_i = s_i^{(1)}, \qquad \sigma_i^2 = s_i^{(2)} - \mu_i^2, \tag{62}$$

où

$$s_i^{(\alpha)} = \frac{\sum_{m=1}^{M}\sum_{n=1}^{N} I_{[\lambda_{m,n}=\psi_i]} X_{m,n}^{\alpha}}{\sum_{m=1}^{M}\sum_{n=1}^{N} I_{[\lambda_{m,n}=\psi_i]}}, \quad \alpha=1,2. \tag{63}$$

On remarquera que, à quelques opérations de comptage et moyennage près, la phase d'apprentissage ne requiert, à ce stade, aucun calcul qui ne soit nécessaire à la phase de reconnaissance. Le lecteur vérifiera aisément qu'il n'en irait plus de même si la ré-estimation des probabilités de transition était fondée sur une ré-écriture de (53) exploitant une approximation du même type que (60) pour la probabilité *a posteriori* d'effectuer la transition $\psi_i - \psi_j$ en position $(m, n)-(m, n+1)$. En effet, cette méthode demanderait la détermination de la fréquence avec laquelle chaque transition possible est la plus probable *a posteriori* : C'est une question dont nous ne nous sommes pas souciés précédemment.

Par souci d'efficacité, nous introduirons comme réestimateur de probabilité de transition la fréquence empirique de la réalisation de chaque transition possible au cours d'une phase de reconnaissance. A défaut de cohérence au niveau conceptuel, nous obtenons de cette manière un algorithme tout à fait cohérent du point de vue opérationnel. En effet, la formule de réestimation de P_{ij} s'écrit :

$$P_{ij} = \frac{\sum_{m=1}^{M}\sum_{n=1}^{N-1} I_{[\lambda_{m,n}=\psi_i \wedge \lambda_{m,n+1}=\psi_j]} + \sum_{n=1}^{N}\sum_{m=1}^{M-1} I_{[\lambda_{m,n}=\psi_i \wedge \lambda_{m+1,n}=\psi_j]}}{2\sum_{m=1}^{M-1}\sum_{n=1}^{N-1} I_{[\lambda_{m,n}=\psi_i]} + \sum_{m=1}^{M-1} I_{[\lambda_{m,N}=\psi_i]} + \sum_{n=1}^{N-1} I_{[\lambda_{M,n}=\psi_i]}}, \tag{64}$$

ce qui concorde parfaitement avec (61)-(63). (Le facteur 2 apparaissant au dénominateur trouve son origine dans le fait que tout pixel n'appartenant pas à la dernière ligne ou à la dernière colonne sert simultanément d'origine à une transition horizontale et une transition verticale.)

Arrivé à ce stade d'approximation par rapport à notre modèle de départ, on pourrait difficilement justifier le recours à une traduction *fréquentielle* de (57) et (58) dans le cas de champs de Pickard multi-valués. En fait, nous verrons à la Section 5 que l'utilisation systématique de (64) se justifie pleinement.

Dans le cas de variables aléatoires indépendantes et identiquement distribuées, il est bien connu que du point de vue théorique l'algorithme DD souffre de tares rédhibitoires. Ainsi, il se caractérise par une tendance à sous-estimer les écarts types et à surestimer les distances entre les moyennes de distributions empiétantes. En fait, on a montré que, d'une manière générale, cette méthode fournit des résultats non consistants et asymptotiquement biaisés [9]. On peut difficilement espérer un comportement plus sain dans le cas markovien qui nous occupe. Toutefois, nous allons voir ci-dessous qu'à défaut de susciter l'estime des théoriciens la technique DD est plus que digne d'intérêt de la part des praticiens.

5. Résultats expérimentaux

Les algorithmes d'apprentissage et de reconnaissance formulés ci-dessus ont été traduits dans le langage de programmation PASCAL et de nombreux essais ont été exécutés dans le but d'étayer la validité du modèle proposé et d'évaluer l'efficacité de l'approche que nous avons élaborée. Les résultats obtenus avec quatre images à niveaux de gris — une artificielle et trois réelles — font l'objet de la discussion qui suit. Les problèmes traités relèvent de la restauration d'image dans le premier cas, de la segmentation dans les trois autres.

Conformément aux considérations émises à la section 4.1, nous définirons dans chaque cas une famille de modèles par un minimum de spécifications *a priori* et nous rechercherons, parmi l'ensemble des modèles, celui qui s'accorde le mieux aux données expérimentales. Plus précisément, nous nous contenterons de spécifier le nombre d'états souhaités et de supposer que les distributions conditionnelles des niveaux de gris observés sont approximées par des lois gaussiennes.

Dans la version de notre programme orientée vers la segmentation des images à niveaux de gris en représentation d'un octet par pixel, une fois fixé le nombre d'états, la détermination des valeurs initiales des paramètres à estimer s'opère de manière automatique :

(i) $P_i = 1/\vartheta$, $P_{ii} = 0{,}5$, $\forall i$, $P_{ij} = 1/2(\vartheta-1)$, $\forall i, j$, $i \neq j$; pour $\vartheta = 2$, nous supposons l'équi-probabilité la plus parfaite; pour $\vartheta > 2$, nous supposons l'image constituée de régions uniformes d'une « certaine » étendue.

(ii) On assigne à μ_i la valeur du quantile de $(i/(\vartheta+1))$ de la fonction empirique de répartition — ou histogramme cumulé — des niveaux de gris, $i=1,\ldots,\vartheta$. Ce choix garantit une répartition relativement uniforme et couvrant toute la dynamique de l'image.

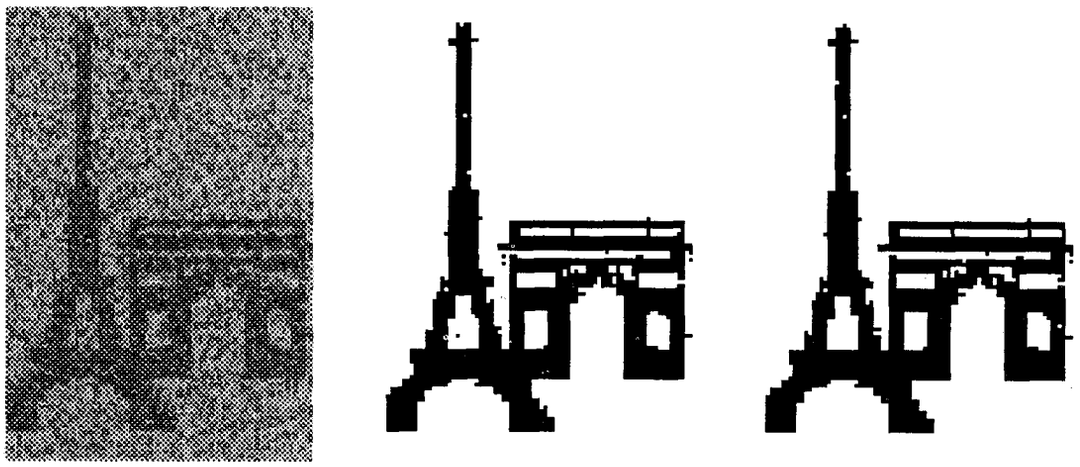


Fig. 4. — Résultats de restauration de l'image initiale (*a*) après apprentissage de type EM (*b*) et DD (*c*).

(iii) $\sigma_i=(\mu-\mu_1)/2(\vartheta-1)$, $i=1,\ldots,\vartheta$; les valeurs initiales sont supposées être séparées, en moyenne, par une distance de deux écarts types.

Notre première illustration a trait à une image artificielle, de taille 120×80, qui est une version dégradée (bruitée) d'un poster qui était très populaire, fin 1986, dans les milieux français de la reconnaissance des formes. A l'image binaire de départ — préalablement convertie en image à deux niveaux de gris (de valeurs 100 et 140) — est ajoutée une composante de bruit blanc, de moyenne nulle et d'écart type égal à 20. L'image ainsi obtenue que nous nous proposons de restaurer est représentée à la figure 4 (*a*).

Dans le cas qui nous occupe, il était naturel de choisir un modèle à deux états possibles, et les figures 4 (*b*) et (*c*) montrent les résultats de restauration obtenus avec les paramètres estimés par les algorithmes d'apprentissage de type EM et DD respectivement. A l'évidence, ces deux résultats sont très semblables. Pour s'en convaincre, il suffit de se rendre compte qu'il serait impossible de décider lequel des deux procède de l'apprentissage le plus élaboré.

L'expérience nous a montré que dans le voisinage d'un optimum (local), la restauration est pratiquement insensible à de faibles perturbations de l'ensemble des paramètres. Cette observation n'est pas sans importance en ce qu'elle nous permet de fixer *a priori* le nombre souhaitable d'itérations d'apprentissage sans nous soucier de contrôler la convergence de manière numérique. Comme on pouvait s'y attendre, il apparaît que le nombre minimal d'itérations nécessaires varie en raison inverse du rapport signal-bruit de l'image. Dans le cas de la figure 4 (*a*), pour laquelle ce rapport est particulièrement faible, l'expérience montre que, pour l'algorithme EM, il est inutile de poursuivre l'apprentissage au-delà de la vingtième itération. Cela n'apporte plus d'amélioration à la restauration obtenue ([8]). Des nombres d'itérations nettement plus faibles donnent de bons résultats sur des images plus réalistes, c'est-à-dire, moins bruitées.

Maintes expériences antérieures avec des images artificielles nous ont appris que la fréquence d'erreur de l'estimation de l'état des pixels n'est pas un indicateur satisfaisant — du point de vue de la perception — de la qualité d'une restauration. Nous y ferons cependant appel ici dans le but de comparer la vitesse de convergence des deux algorithmes d'apprentissage. Au cours d'une série d'expériences, nous avons exécuté l'algorithme de reconnaissance avec les paramètres estimés à l'issue de la n-ième itération ($n=1,\ldots,15$) de chacun des algorithmes d'apprentissage, et nous avons déterminé le pourcentage d'erreur de chaque restauration par rapport à l'image (non bruitée) initiale. Ces pourcentages sont illustrés à la figure 5.

Remarquons tout d'abord que, les paramètres des modèles étant connus, la probabilité d'erreur théori-

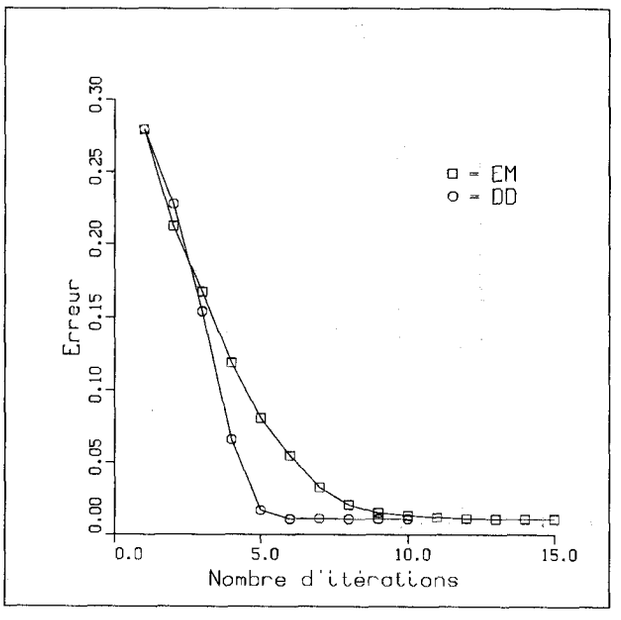


Fig. 5. — Comparaison des vitesses de convergence des algorithmes d'apprentissage.

([8]) Du point de vue numérique, la convergence vers les paramètres (connus) du modèle est remarquable : à la dernière itération de l'algorithme d'apprentissage, les valeurs estimées des moyennes et écarts types sont de 100,34, 140,24, 20,51 et 19,82 respectivement.

que associée à une règle bayesienne ne prenant pas en compte l'information contextuelle que traduit la dépendance markovienne peut être déterminée : elle est de l'ordre de 15 %. On peut voir à la figure 5 que notre méthode de choix de valeurs initiales des paramètres à estimer entraîne une fréquence d'erreur de 28 %, et que les deux méthodes d'apprentissage permettent de réduire cette fréquence à une valeur très satisfaisante de 1 %. La seule différence notable entre les deux méthodes d'apprentissage est leur vitesse de convergence, celle de l'algorithme DD étant pratiquement le double de celle de l'algorithme EM.

Les trois illustrations suivantes font appel à des images réelles à niveaux de gris (0, . . ., 255), de taille 128×128. Toutes trois ont pour caractéristique commune d'être relativement pauvres en composantes hautes fréquences ce qui permet d'utiliser notre modèle à des fins de segmentation. Avec la plupart des images de ce genre nous avons constaté expérimentalement qu'en général six états distincts suffisent à extraire l'information signifiante, et ce nombre d'états a été utilisé dans chacune des expériences décrites ci-dessous ([9]). De plus, pour chacun de ces trois cas, le nombre d'itérations d'apprentissage a été fixé arbitrairement à dix.

Notre seconde illustration est représentée à la figure 6 (*a*). L'histogramme de cette image est un mélange compliqué de modes et de plateaux. Certaines parties de l'objet qui nous intéresse, l'avion, sont faiblement contrastées vis-à-vis d'un fond irrégulier. En toute rigueur, l'application de notre technique à un problème dans lequel l'espace d'état est multivalué aurait dû nous conduire, dans le cas de l'apprentissage de type EM, à utiliser les formules (57) et (58) pour satisfaire la condition de réversibilité de la chaîne engendrée par les probabilités de transition estimées. Toutefois, des essais préalables au moyen d'un programme mettant en œuvre la formule (52) ont montré que les matrices de probabilités de transition estimées sont en fait, avec une très bonne approximation, naturellement symétriques. Un moment de réflexion per-

([9]) On notera la prolifération rapide des paramètres à estimer en fonction du nombre d'états : dans le cas présent, pas moins de 57 paramètres doivent être estimés.

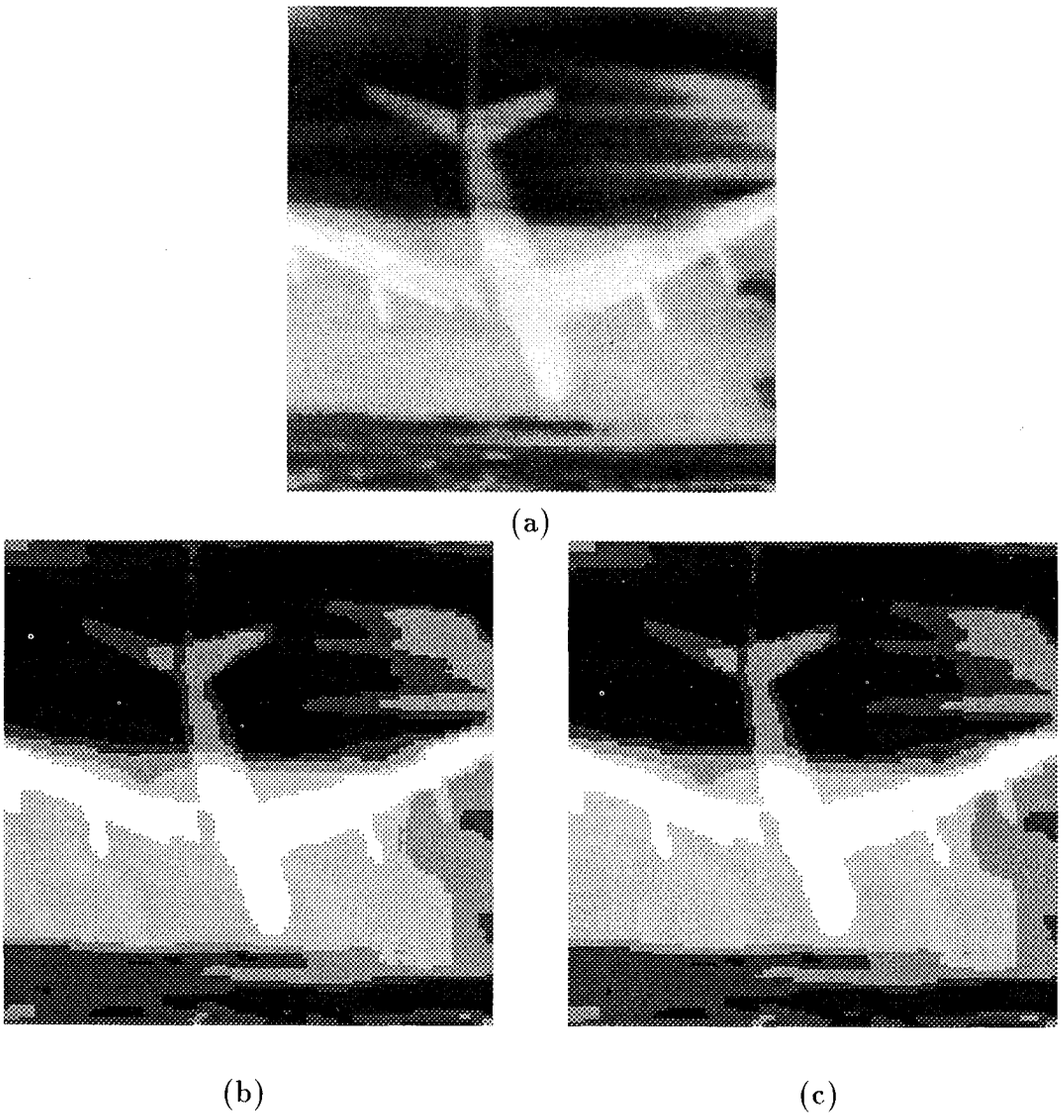


Fig. 6. — Résultats du traitement de l'image originale (*a*) avec apprentissage de type EM (*b*) et DD (*c*).

met de s'assurer qu'il est normal qu'il en soit ainsi dans le cas des images à niveaux de gris modérément contrastées. En tout état de cause, l'algorithme EM établi pour le champ binaire a été utilisé indépendamment du nombre d'états souhaité et a donné des résultats tout à fait satisfaisants, tandis que l'algorithme DD a été utilisé tel que décrit à la section 4.3.2.

Afin d'améliorer le rendu du contraste apparent des segmentations obtenues, les niveaux de gris ont été répartis — par interpolation linéaire — sur toute la plage dynamique. La figure 6(*b*) représente la segmentation obtenue à l'issue d'un apprentissage de type EM tandis que la figure 6(*c*) représente la segmentation obtenue à l'issue d'un apprentissage de type DD. A nouveau, les deux types d'apprentissage proposés font montre d'une concordance remarquable. De plus, il est particulièrement satisfaisant de constater que la segmentation obtenue est en tout point conforme à notre attente subjective en ce que les ailes et le fuselage d'une part, la queue de l'appareil d'autre part peuvent être extraits de l'image sous la forme de composantes connexes ([10]).

Nos deux dernières illustrations, montrent les résultats de segmentation obtenus après apprentissage de type DD uniquement. Aux figures 7 et 8, on peut voir en (*a*) les images initiales, en (*b*) les images à six niveaux obtenues après segmentation et répartition du contraste sur toute la plage dynamique des niveaux de gris, et en (*c*) les images de bords correspondantes ([11]). Le résultat obtenu dans le cas de la figure 7 est particulièrement satisfaisant. En particulier, les régions sont à ce point

([10]) Dans le cas de cette image, la segmentation obtenue est pratiquement stabilisée dès la cinquième itération des algorithmes d'apprentissage.

([11]) Ces images de bords sont obtenues en plaçant un élément de bord (horizontal ou vertical) entre toute paire de pixels de niveaux de gris différents. Par conséquent, ces images sont représentées sur une grille de taille 257×257. Pour plus de détails, *voir* [10].

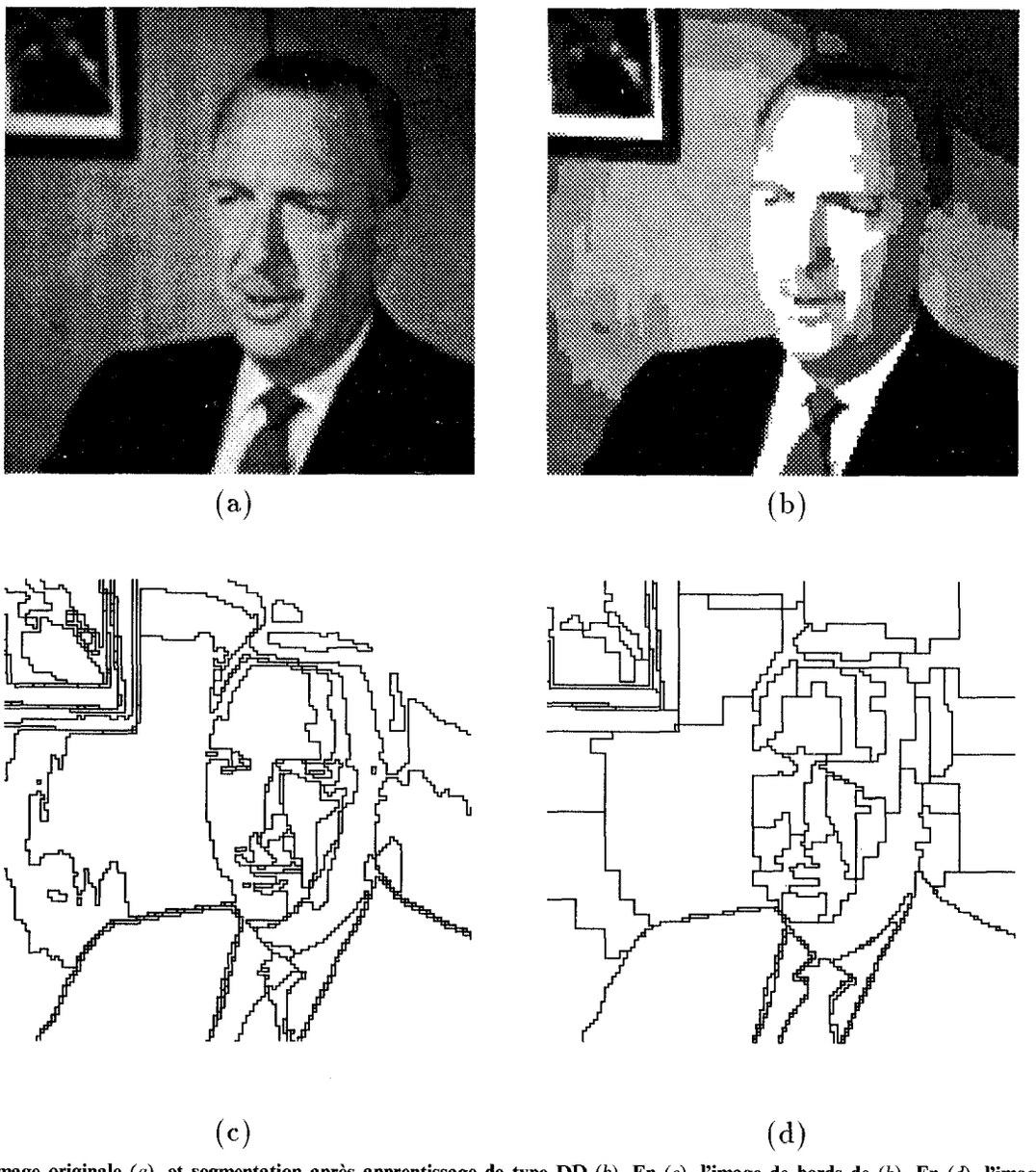
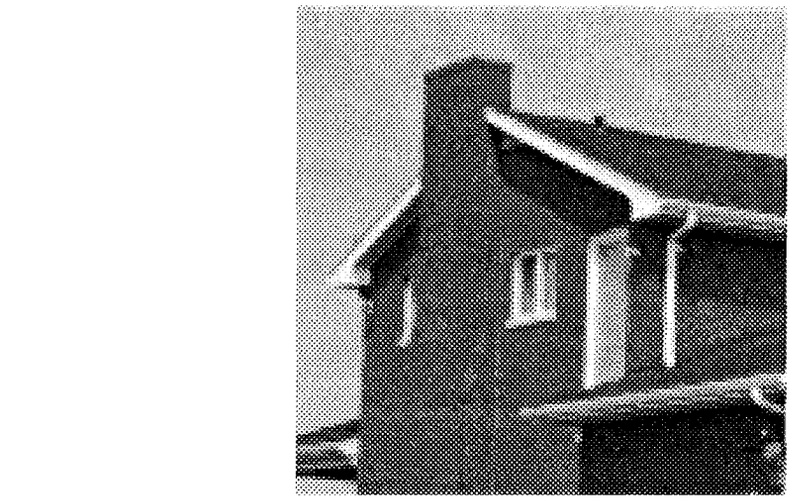
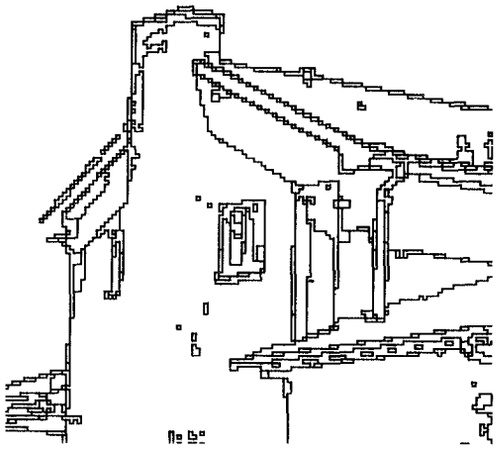


Fig. 7. — Image originale (*a*), et segmentation après apprentissage de type DD (*b*). En (*c*), l'image de bords de (*b*). En (*d*), l'image de bords obtenue après segmentation par l'algorithme « Split and Merge ».



(a)

(b)



(c)

Fig. 8. – Résultats du traitement de l'image originale (*a*) avec apprentissage de type DD (*b*). En (*c*) l'image de bords de (*b*).

« significatives » qu'il est possible de reconnaître un visage à partir de l'image de bords 7 (*c*) seulement.

A titre de comparaison, la figure 7 (*d*) illustre la segmentation obtenue au moyen de l'algorithme « Split and Merge » de Horowitz et Pavlidis [33, 45]. Cet algorithme qui est fondé sur la représentation de l'image en arbre quaternaire a une tendance manifeste à produire des régions rectangulaires et l'interprétation de l'image de bords devient bien plus malaisée ([12]). Nous avons pu observer le même phénomène dans nombre d'autres cas et, en particulier, dans celui de l'avion de la figure 6.

L'image de départ utilisée pour notre dernière illustration à la figure 8 (*a*) est un autre « classique » du traitement d'images. On peut voir que toutes les zones de gris, même celles qui ne sont que très faiblement contrastées sont correctement identifiables dans l'image de bords 8 (*c*). D'une manière générale, il est fréquent de faire suivre une segmentation par une passe d'élimination des régions d'étendue faible qui, le plus souvent, ne sont pas significatives. [On retrouve cette idée dans la version disponible dans la librairie SPIDER de l'implantation de la procédure Split and Merge qui a servi à produire la segmentation de la figure 7 (*d*).] Nous n'avons pas eu recours à cette technique dans le cas présent par souci de fidélité au modèle markovien que nous nous proposions d'illustrer. Il est cependant évident que l'élimination des « petites » régions de la figure 8 (*c*) produirait une segmentation d'autant plus significative.

([12]) Pour s'en convaincre, il suffit de masquer les deux images de bords à hauteur des épaules de Walter Cronkite.

6. Conclusions

Dans cet article nous avons exploré un certain nombre des possibilités offertes par une modélisation des images digitales qui fait appel aux champs aléatoires markoviens de Pickard. Après avoir situé le champ de Pickard dans le contexte plus général des champs aléatoires markoviens, nous avons montré qu'une

méthode de reconnaissance développée par Haslett, et exploitant les structures de chaîne de Markov présentes dans les champs de Pickard, se ramène à deux exécutions orthogonales de l'algorithme *Forward Backward*, un des outils de prédilection des praticiens de la reconnaissance de la parole. Nous avons proposé deux algorithmes qui permettent d'estimer de manière non supervisée les paramètres du modèle bi-dimensionnel proposé. Nous avons aussi démontré au moyen de quelques exemples que notre technique permet de traiter avec succès les problèmes de restauration et de segmentation d'images digitales à niveaux de gris.

Une étude comparative de divers algorithmes de segmentation que nous nous proposons de présenter ailleurs fait clairement apparaître que :

(i) la méthode présentée ici est en mesure de rivaliser avec les techniques connues réputées les plus performantes;

(ii) tout en étant d'une facilité d'usage exceptionnelle.

En effet, la plupart des algorithmes disponibles actuellement imposent au praticien le choix préalable de nombre de paramètres : définition du contraste inter-régions minimal, écart maximal admissible entre niveaux de gris au sein d'une même région, tailles extrêmes des régions, seuil de démarrage de l'opérateur phagocyte, etc. Le choix de ces paramètres n'est pas sans avoir une importance majeure à la fois sur la qualité subjective de la segmentation obtenue et sur le temps de calcul nécessaire à la bonne terminaison de l'algorithme en question. Par contre, l'activation de notre méthode ne demande que la spécification du nombre souhaité d'états du champ markovien, tous les autres paramètres étant initialisés de manière automatique. Comme nous l'avons signalé plus haut, le nombre « magique » de six états nous a donné systématiquement des résultats plus que satisfaisants sur une très large gamme d'images à niveaux de gris.

Manuscrit reçu le 18 *février* 1988.

BIBLIOGRAPHIE

[1] K. Abend, Compound decision procedures for unknown distributions and for dependent states of nature, in *Pattern Recognition*, L. N. Kanal éd., Washington DC, Thompson Book Cy, 1968, p. 207-249.

[2] K. Abend, T. J. Harley et L. N. Kanal, Classification of binary random patterns, *IEEE Trans. Inform. Theory*, IT-11, Oct. 1965, p. 538-544.

[3] M. Askar et H. Derin, A recursive algorithm for the Bayes solution of the smoothing problem, *IEEE Transactions on Automatic Control*, AC-26, 1981, p. 558-560.

[4] L. E. Baum, An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes, *Inequalities*, 3, 1972, p. 1-8.

[5] L. E. Baum, T. Petrie, G. Soules et N. Weiss, A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains, *Ann. Math. Statist.*, 41, 1972, p. 164-171.

[6] J. Besag, Nearest-neighbour systems and the autologistic model or binary data, *Jl. Royal Stat. Soc*, B-34, 1972, p. 75-83.

[7] J. Besag, Spatial interaction and the statistical analysis of lattice systems, *Jl. Royal Stat. Soc.*, B-36, 1974, p. 192-236.

[8] J. Besag, On the statistical analysis of dirty pictures, *Jl. Royal Stat. Soc.*, B-48, 1985, p. 259-302.

[9] H. Bourlard, C. Wellekens et H. Ney, Connected digit recognition using vector quantization, *Proc. Intern. Conf. Acoustics, Speech and Signal Processing*, San Diego, 1984.

[10] C. R. Brice et C. L. Fennema, Scene analysis using regions, *Artificial Intelligence*, 1, 1970, p. 205-226.

[11] P. Bryant et J. W. Williamson, Asymptotic behavior of classification maximum likelihood estimates, *Biometrika*, 65, 1978, p. 273-281.

[12] P. Carnevalli, L. Coletti et S. Patarnello, Image processing by simulated annealing, *IBM Jl. of Research and Development*, 29, Nov. 1985, p. 569-579.

[13] H. Cerf-Danon, A.-M. Derouault, M. El-Beze, B. Merialdo et S. Soudoplatoff, Speech recognition experiment with 10,000 words dictionary, in *Pattern Recognition Theory and Applications*, P. A. Devijver et J. Kittler éd., Heidelberg, Springer-Verlag, 1987.

[14] G. R. Cross et A. K. Jain, Markov random field texture models, *IEEE Trans. Pattern Anal., Machine Intell.*, PAMI-5, Jan. 1983, p. 24-39.

[15] H. Derin, H. Elliot, R. Christi et D. Geman, Bayes smoothing algorithms for segmentation of binary images modeled by Markov random fields, *IEEE Trans. Pattern Anal., Machine Intell.*, PAMI-6, Nov. 1984, p. 707-720.

[16] P. A. Devijver, Classification in Markov chains for minimum symbol error rate, *Proc. 7th Intern. Conf. Pattern Recognition*, Montreal, 1984, p. 1334-1336.

[17] P. A. Devijver, Probabilistic labeling in a hidden second order Markov mesh, in *Pattern Recognition in Practice II*, E. Gelsema et L. N. Kanal éd., Amsterdam, North Holland, 1985, p. 113-123.

[18] P. A. Devijver, Baum's forward-backward algorithm revisited, *Pattern Recognition Letters*, 3, Dec. 1985, p. 369-373.

[19] P. A. Devijver, Cluster analysis by mixture identification, in *Data Analysis in Astronomy*, V. Di Gesù *et al.* éd., New York, Plenum, 1985, p. 29-44.

[20] P. A. Devijver, Segmentation of binary images using third order Markov mesh image models, in *Proc. 8th Internat. Conf. Pattern Recognition*, Paris, Oct. 1986, p. 259-261.

[21] P. A. Devijver et J. Kittler, *Pattern Recognition: A Statistical Approach*, Englewood-Cliffs, Prentice Hall, 1982.

[22] P. A. Devijver et M. M. Dekesel, Algorithmes d'apprentissage d'un modèle Markovien d'image, *Actes du 6e Congrès RFIA*, Antibes, Nov. 1987, p. 193-207.

[23] P. A. Devijver et M. Dekesel, Learning the parameters of a hidden Markov random field image model: A simple example, in *Pattern Recognition Theory and Applications*, P. A. Devijver et J. Kittler éd., Heidelberg, Springer-Verlag, 1987, p. 141-163.

[24] R. O. Duda et P. E. Hart, *Pattern Classification and Scene Analysis*, New York, Wiley, 1973.

[25] G. D. Forney Jr., The Viterbi algorithm, *Proc. IEEE*, 61, Mar. 1973, p. 268-278.

[26] S. Geman et D. Geman, Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images, *IEEE Trans. Pattern Anal., Machine Intell.*, PAMI-6, Nov. 1984, p. 721-741.

[27] D. GEMAN, S. GEMAN et C. GRAFFIGNE, Locating texture and object boundaries, in *Pattern Recognition Theory and Applications*, P. A. DEVIJVER et J. KITTLER éd., Heidelberg, Springer-Verlag, 1987.

[28] S. GÜLER, G. GARCIA, L. GÜLEN et M. N. TOKSÖZ, The detection of geological fault lines in radar images, in *Pattern Recognition Theory and Applications*, P. A. DEVIJVER et J. KITTLER éd., Heidelberg, Springer-Verlag, 1987.

[29] A. R. HANSON, E. M. RISEMAN et E. FISHER, Context in word-recognition, *Pattern Recognition*, 8, 1976, p. 35-46.

[30] R. M. HARALICK, Decision making in context, *IEEE Trans. Pattern Anal., Machine Intell.*, PAMI-5, July 1983, p. 417-428.

[31] J. HASLETT, Maximum likelihood discriminant analysis on the plane using a Markovian model of spatial context, *Pattern Recognition*, 18, 1985, p. 287-296.

[32] M. HASSNER et J. SKLANSKY, The use of Markov random fields as models of texture, *CGIP*, 12, 1980, p. 357-370.

[33] S. L. HOROWITZ et T. PAVLIDIS, Picture segmentation by a tree traversal algorithm, *JACM*, 23, 1976, p. 368-388.

[34] F. JELINEK et R. L. MERCER, Interpolated estimation of Markov source parameters from sparse data, in *Pattern Recognition in Practice*, E. GELSEMA et L. KANAL éd., Amsterdam, North-Holland, 1980, p. 381-397.

[35] F. JELINEK, R. L. MERCER et L. R. BAHL, Continuous speech recognition: Statistical methods, in *Handbook of Statistics* 2, P. R. KRISHNAIAH et L. N. KANAL éd., Amsterdam, North-Holland, 1982, p. 549-573.

[36] L. N. KANAL, Markov mesh models, *Computer Graphics and Image Processing*, 12, 1980, p. 371-375, (also in *Image Modeling*, A. ROSENFELD éd., New York, Academic Press, 1981, p. 239-243).

[37] J. G. KEMENI et J. L. SNELL, *Finite Markov Chains*, New York, Springer-Verlag, 1976.

[38] R. KINDERMAN et J. L. SNELL, *Markov Random Fields and their Applications*, Providence RI, American Mathematical Society, 1980.

[39] V. LACROIX, Pixel labeling in a second-order Markov mesh, *Signal Processing*, 12, Jan. 1987, p. 59-82.

[40] D. S. LEBEDEV, Probabilistic characterization of images in filtration and restoration problems, in *Signal Processing: Theories and Applications*, M. KUNT et F. DE COULON éd., Amsterdam, North-Holland, 1980, p. 55-64.

[41] S. E. LEVINSON, Continuously variable duration hidden Markov models for automatic speech recognition, *Computer, Speech and Language*, 1, 1986, p. 29-45.

[42] S. E. LEVINSON, L. R. RABINER et M. M. SONDHI, An introduction to the application of the theory of probabilistic functions of a Markov process in automatic speech recognition, *BSTJ*, 62, 1983, p. 1035-1074.

[43] L. A. LIPORACE, Maximum likelihood estimation for multivariate observations of Markov sources, *IEEE Trans. Inform. Theory*, IT-28, 1982, p. 729-734.

[44] D. L. NEUHOF, The Viterbi algorithm as an aid to text-recognition, *IEEE Trans. Inform. Theory*, IT-21, Mar. 1975, p. 222-226.

[45] T. PAVLIDIS, *Structural Pattern Recognition*, Berlin, Springer-Verlag, 1980.

[46] D. K. PICKARD, A curious binary lattice process, *Jl. Appl. Prob.*, 14, 1977, p. 717-731.

[47] D. K. PICKARD, Unilateral Markov fields, *Adv. Applied Probability*, 12, 1982, p. 655-671.

[48] L. R. RABINER, S. E. LEVINSON et M. M. SONDHI, On the application of vector quantization and hidden Markov models to speaker independent, isolated word recognition, *BSTJ*, 62, 1983, p. 1075-1105.

[49] J. RAVIV, Decision making in Markov chains applied to the problem of pattern recognition, *IEEE Trans. Inform. Theory*, IT-3, 1967, p. 536-551.

[50] R. A. REDNER et H. F. WALKER, Mixture densities, maximum likelihood and the EM algorithm, *SIAM Review*, 26, 1984, p. 195-239.

[51] E. RISEMAN et R. W. EHRICH, Contextual word recognition using binary digrams, *IEEE Trans. Comput.*, C-20, April 1971, p. 397-403.

[52] R. SHINGHAL, D. ROSENBERG et G. T. TOUSSAINT, A simplified heuristic version of a recursive Bayes algorithm for using context in text recognition, *IEEE Trans. Syst., Man, Cybern.*, SMC-8, May 1978, p. 412-414.

[53] G. T. TOUSSAINT, The use of context in pattern recognition, *Pattern Recognition*, 10, 1978, p. 189-204.

[54] G. WOLBERG et T. PAVLIDIS, Restoration of binary images using stochastic relaxation with annealing, *Pattern Recognition Letters*, 3, 1985, p. 375-388.

[55] S. YAKOVITZ, Unsupervised learning and the identification of finite mixtures, *IEEE Trans. Inform. Theory*, IT-16, 1970, p. 330-338.

[56] S. YAKOVITZ et J. SPRAGINS, On the identifiability of finite mixtures, *Ann. Math. Stat.*, 39, 1968, p. 209-214.