

Modèles de Markov en Imagerie

Wojciech Pieczynski
Institut National des Télécommunications
Département CITI
9, rue Charles Fourier, 91000 Evry

Résumé

L'objet de l'article est de présenter quelques aspects des traitements statistiques des images utilisant des modèles de Markov. En choisissant pour cadre la segmentation statistique nous rappelons brièvement la nature et l'intérêt des traitements probabilistes et présentons les modèles de Markov cachés classiques : champs, chaînes, et arbres. Les méthodes Bayésiennes de segmentation MPM et MAP sont décrites, ainsi que les méthodes d'apprentissage EM et ICE. Nous terminons par la description de quelques modèles ou méthodes de traitements plus récents comme les modèles de Markov Couple, la fusion de Dempster-Shafer dans le contexte Markovien, ou l'estimation des mélanges généralisés.

Abstract

The aim of this paper is to present some aspects of Markov model based statistical image processing. Within the context of image segmentation, we first review the interest of « pixel by pixel » statistical segmentation, and then proceed to present classical hidden Markov models, including fields, chains, and trees. Classical Bayesian segmentation methods such as MPM and MAP, and parameter estimation techniques like EM or ICE, are then described. Finally, we mention some recent models and processing techniques, including Pairwise Markov models, generalized mixture estimation, and Dempster-Shafer fusion in the Markov models context.

1. Introduction

L'objet de cet article est de préciser l'intérêt que peuvent présenter les modélisations probabilistes et les traitements statistiques en imagerie, avec un accent particulier mis sur l'utilisation des modèles Markoviens. Le cadre illustratif choisi est celui de la segmentation. Ce problème est simple dans sa formulation et important dans les applications, ce qui nous semble bien adapté pour exposer l'intérêt profond des méthodes de traitement probabilistes.

La nature de la classification Bayésienne et son intérêt en segmentation d'image est présentée, de manière intuitive et en dehors des modèles de Markov, dans la section suivante. La section 3 contient une description des modèles de Markov cachés classiques : champs, chaînes, et arbres. La section 4 est consacrée à un modèle récent dit champs de Markov Couple. L'importante question de l'estimation des paramètres, permettant de rendre les traitements automatiques, est abordée dans la section 5 avec la description, en particulier, des méthodes EM et ICE. Diverses possibilités d'extensions aux modèles plus complexes sont discutées dans la section 6 et la dernière section 7 contient les conclusions et perspectives.

2. Modèle probabiliste et classification Bayésienne

Les modélisations probabilistes et les traitements statistiques peuvent s'avérer d'une remarquable efficacité dans le traitement de certains problèmes se posant en imagerie. De manière très générale, on peut envisager de faire appel à des méthodes probabilistes de traitement lorsque l'on se trouve dans une situation réunissant les deux points suivants : (i) on souhaite déterminer des caractéristiques de l'image, non observables directement et représentées par un élément x d'un espace X , à partir des caractéristiques de l'image observables et représentées par un élément y d'un espace Y ; (ii) il n'existe manifestement pas de liens déterministes entre les y et les x (en d'autres termes, plusieurs x possibles peuvent correspondre à un y observé). Le deuxième point semble interdire définitivement tout traitement sérieux du problème ; cependant, un traitement statistique est souvent possible. De plus, ce type de traitement peut bénéficier de trois qualités majeures : généralité, optimalité, et souplesse.

Précisons les sens donnés à ces trois qualités dans le contexte considéré.

Il n'existe donc pas de liens déterministes entre les y et les x ; cependant, l'observation de y doit apporter une certaine information sur x . Dans le contexte qui nous intéresse, cette information est

donnée par une loi de probabilité sur X et peut être appréhendée, par la loi de grands nombres, de la façon suivante : si la situation se reproduit un certain nombre de fois, la probabilité $p(A|y)$ d'un ensemble $A \subset X$ est la proportion des x se trouvant dans A . La « généralité » des traitements statistiques est alors assurée par les diverses possibilités de définir les lois de probabilités, conditionnelles à y , sur X . A y fixé, cette « généralité » se traduit par la possibilité de modélisation des situations extrêmement diverses, allant du lien déterministe à l'indépendance totale entre y et x . En effet, dans le premier cas la probabilité $p(\cdot|y)$ est une masse de Dirac et dans le deuxième cas cette probabilité ne dépend pas de y . Il est ainsi possible de modéliser des liens « plus au moins » forts entre y et x , de plus, la situation où les liens entre y et x sont déterministes apparaît comme un cas particulier des modélisations probabilistes. Concernant l'optimalité considérons une fonction $L: X^2 \rightarrow R^+$ modélisant la gravité des erreurs : $L(x^1, x^2)$ représente le coût de l'erreur « on pense que x est x^1 , alors que le vrai x est x^2 ». Notons que la fonction L , qui est appelée « fonction de perte », ou encore « fonction de coût », est indépendante des modélisations probabilistes éventuellement retenues. Supposons que l'on dispose d'une règle de décision $d: Y \rightarrow X$ et l'on souhaite évaluer sa qualité. Connaissant la probabilité $p(\cdot|y)$, on peut calculer le « risque » défini par

$$R(L, d, y) = \int_x L(d(y), x) p(x|y) dx \quad (2.1)$$

Notons que le « risque » est une quantité très « parlante » pour un praticien. En effet, $R(L, d, y)$ désigne la « perte moyenne » : si l'on utilise la règle d un nombre n suffisamment grand de fois, la perte globale subie est le produit de n par $R(L, d, y)$. A y observé, on peut alors rechercher $d_B^L(y) \in X$ qui minimise l'intégrale ($d_B^L(y)$ est dite « règle Bayésienne » liée à L), et qui minimisera la perte globale. Il reste à préciser une dernière généralisation. Lorsque l'on est confronté à n expériences, les n observations y peuvent ne pas être toutes égales et évoluer de manière aléatoire en suivant une loi de probabilité définie sur Y . Précisons dès à présent que cette nouvelle généralisation ne modifie pas la solution optimale d_B^L , qui reste celle qui minimise la perte globale. Cependant, préciser la manière dont on prend en compte l'aspect aléatoire de y nous permettra de

terminer la modélisation probabiliste du problème. Les quantités x et y étant aléatoires, il est pratique de les considérer comme réalisations des variables aléatoires X et Y . La loi $p(x, y)$ du couple (X, Y) est donnée par la loi $p(y)$ de Y et les lois $p(x|y)$ de X conditionnelle à Y . Le risque moyen associé à L et d est alors défini par

$$R(L, d) = E[L(d(Y), X)] \quad (2.2)$$

et la décision Bayésienne associée à L est la décision qui minimise ce risque moyen :

$$E[L(d_B^L(Y), X)] = \min_d E[L(d(Y), X)] \quad (2.3)$$

Finalement, la qualité d'optimalité peut se résumer de la manière suivante. Si la loi $p(x, y)$ du couple (X, Y) est connue, la règle de décision d_B^L qui consiste à minimiser, pour chaque y observé, le risque défini par (2.1), est celle parmi toutes les règles possibles, d'origine probabiliste ou pas, qui minimise la perte globale subie lorsque l'on traite un nombre suffisamment grand des y . De plus, le calcul de $E[L(d_B^L(Y), X)]$ nous donne le montant de la perte minimale.

Enfin, les méthodes probabilistes de traitement présentent une « souplesse » dans le sens où l'optimalité discutée ci-dessus est adaptable à des préoccupations particulières. En effet, devant une réalité « objective » donnée, modélisée par la loi $p(x, y)$ du couple (X, Y) , deux clients peuvent avoir des préoccupations différentes – voire contradictoires – modélisées par deux fonctions de perte différentes L^1, L^2 , qui donnent le montant de l'argent perdu, à long terme, par chacun des clients. Ces fonctions vont définir deux règles de décisions différentes d^1, d^2 . Ainsi d^1 est optimale pour le premier client, qui perdrait plus d'argent en appliquant d^2 , et d^2 est optimale pour le deuxième client, qui perdrait plus d'argent en appliquant d^1 .

Exemple 2.1

Considérons le problème d'établissement d'une carte d'une région comportant deux classes « eau », notée E , et « forêt », notée F . Supposons que l'on souhaite travailler « pixel par pixel », ce qui signifie que sur chaque pixel on observe un niveau de gris y et l'on doit lui attribuer un x dans l'ensemble $X = \{E, F\}$. On se trouve dans un contexte probabiliste si un niveau de gris donné peut aussi bien correspondre à de l'eau qu'à de la forêt, ce qui arrive fréquemment dans des images fortement « bruitées », fournies par des capteurs radar par

exemple. La loi $p(x, y)$ du couple (X, Y) peut alors être donnée par la loi $p(x)$ de X , qui est une probabilité sur $X = \{E, F\}$ et désigne simplement la proportion de deux classes dans la région considérée, et par les deux lois $p(y|E)$, $p(y|F)$ de Y conditionnelles à X . Notons que ces deux dernières lois modélisent simultanément la « variabilité naturelle » de deux classes et diverses dégradations dues à l'acquisition, la transmission, et la visualisation des données. Supposons qu'un premier client, pour lequel la surévaluation de la classe « eau » est plus grave que celle de la classe « forêt », perd 1 euro quand un pixel « eau » est classé à tort comme étant « forêt », et 5 euros quand un pixel « forêt » est classé à tort comme étant « eau ». Un deuxième client, pour qui la surévaluation de la classe « forêt » est plus grave que celle de la classe « eau », perd 1 euro quand un pixel « forêt » est classé à tort comme étant « eau », et 10 euros quand un pixel « eau » est classé à tort comme étant « forêt ». Nous avons donc

$$L_1(E, E) = L_1(F, F) = 0,$$

$$L_1(F, E) = 1, L_1(E, F) = 5,$$

et

$$L_2(E, E) = L_2(F, F) = 0,$$

$$L_2(F, E) = 10, L_2(E, F) = 1.$$

Les deux minimisations de (2.1), qui s'écrit ici $L(d(y), E)p(E|y) + L(d(y), F)p(F|y)$, donnent

$$d^1(y) = E \text{ si } p(E|y) \geq 5p(F|y),$$

$$d^1(y) = F \text{ si } p(E|y) \leq 5p(F|y),$$

et

$$d^2(y) = E \text{ si } 10p(E|y) \geq p(F|y),$$

$$d^2(y) = F \text{ si } 10p(E|y) \leq p(F|y)$$

(rappelons que $p(E|y)$ et $p(F|y)$ sont calculées à partir de $p(x)$, $p(y|E)$, et $p(y|F)$ par

$$p(E|y) = \frac{p(E)p(y|E)}{p(E)p(y|E) + p(F)p(y|F)} \quad (2.4)$$

$$p(F|y) = \frac{p(F)p(y|F)}{p(E)p(y|E) + p(F)p(y|F)}.$$

3. Modèles de Markov cachés

3.1 Méthodes locales contextuelles

L'utilisation des décisions Bayésiennes précisées dans le paragraphe précédent peut être mise en défaut lorsque les ensembles X et Y sont de cardinaux importants. En reprenant la situation de l'Exemple 2.1 ci-dessus,

notons S l'ensemble des pixels. Nous avons vu qu'il était possible d'établir une carte de manière optimale en procédant « pixel par pixel ». Pour chaque s dans S , les variables X_s et Y_s sont à valeurs dans $X = \{E, F\}$ et $Y = R$ respectivement. La réalisation $X_s = x_s$ peut être estimée de manière optimale à partir de la réalisation $Y_s = y_s$; cependant, il est légitime d'imaginer que d'autres $Y_t = y_t$ contiennent de l'information utile à la classification du pixel s . En d'autres termes, pour attribuer une classe à s , regarder les niveaux de gris sur s et sur son voisinage devrait être plus efficace que de le regarder uniquement sur s . On peut effectivement montrer que lorsque l'on utilise un voisinage V_s de s , l'utilisation de $Y_{V_s} = (Y_t)_{t \in V_s}$ à la place de Y_s permet d'améliorer l'efficacité des stratégies Bayésiennes : pour toute fonction de perte L , la perte moyenne minimale $E[L(d_B^L(Y_{V_s}), X_s)]$ est inférieure ou égale à la perte moyenne minimale $E[L(d_B^L(Y_s), X_s)]$. De telles méthodes de traitement, dites « locales contextuelles », ont été utilisées et donnent dans certaines situations des résultats de qualité suffisante. Cependant, on se heurte rapidement à des problèmes de calcul lorsque la taille de V croît. Pour deux classes, il est ainsi difficile de prendre en compte des voisinages de taille supérieure à huit. Le nombre de pixels étant couramment de l'ordre de cinquante mille, on est en droit d'imaginer qu'une partie non négligeable de l'information disponible ne peut être exploitée par les méthodes locales.

3.2 Champs de Markov cachés

L'introduction des champs de Markov cachés [6, 16, 29] a permis de contourner, de manière particulièrement élégante, les problèmes de calcul mentionnés dans le paragraphe précédent. Afin de fixer les idées, nous considérerons le problème de la segmentation statistique d'image. Pour S l'ensemble des pixels, on considère deux ensembles de variables aléatoires $X = (X_s)_{s \in S}$ et $Y = (Y_s)_{s \in S}$, appelés « champs aléatoires ». Chaque X_s prend ses valeurs dans un ensemble fini de classes $\Omega = \{\omega_1, \dots, \omega_k\}$ et chaque Y_s prend ses valeurs dans R . Si l'ensemble S contient N pixels, nous avons donc $X = \Omega^N$ et $Y = R^N$, ce qui rend les calculs explicites des décisions Bayésiennes du paragraphe précédent impossibles. Comment manier les processus $X = (X_s)_{s \in S}$ et $Y = (Y_s)_{s \in S}$, qui peuvent comporter, chacun, plus d'un million de composantes ? Cela devient possible en adoptant pour la loi du couple (X, Y) le modèle dit « champ

de Markov caché ». Le champ X est un champ de Markov par rapport à un système de voisinages $V = (V_s)_{s \in S}$, avec la forme des V_s indépendante de s , si elle est de la forme

$$p(x) = \gamma \exp \left[- \sum_{c \in C} \varphi_c(x_c) \right] \quad (3.1)$$

avec C l'ensemble des cliques, une clique étant soit un singleton, soit un ensemble des pixels mutuellement voisins au sens de V . A titre d'exemple, on utilise fréquemment les voisinages composés de quatre plus proches voisins, où les cliques sont les singletons, les ensembles contenant deux pixels voisins horizontalement, et les ensembles contenant deux pixels voisins verticalement.

La formule (3.1) ne permet pas le calcul de $p(x)$ car la constante γ est inconnue ; cependant, et c'est cette propriété qui est fondamentale pour la suite, la loi de X_s conditionnelle à la restriction de X au voisinage V_s est calculable par

$$p(x_s | x_{V_s}) = \frac{\exp \left[- \sum_{c \in C, s \in c} \varphi_c(x_s, x_{\bar{c}}) \right]}{\sum_{\omega \in \Omega} \exp \left[- \sum_{c \in C, s \in c} \varphi_c(\omega, x_{\bar{c}}) \right]} \quad (3.2)$$

où $\bar{c} = c - \{s\}$.

Considérons un $x^0 \in \Omega^N$ quelconque et, en faisant parcourir à s l'ensemble S , dans un balayage « ligne par ligne » par exemple, faisons un tirage aléatoire dans Ω selon la loi (3.2). Dans ces tirages, on tient compte, pour chaque s , des nouvelles valeurs de certains pixels de son voisinage, éventuellement différentes de celles qui composent x^0 . Notons x^1 l'élément de Ω^N (qui est une image de classes, ou une carte) obtenue après le balayage et recommençons, avec x^1 à la place de x^0 , et ainsi de suite On obtient une suite de réalisations $X^1 = x^1, X^2 = x^2, \dots$ du processus aléatoire X^1, X^2, \dots qui est une chaîne de Markov à valeurs dans Ω^N . On montre alors, en utilisant des propriétés élémentaires des chaînes de Markov, que le processus X^1, X^2, \dots converge en loi vers (3.1).

Il est ainsi possible de simuler, et c'est ce fait là qui est à l'origine des possibilités du contournement des problèmes de calcul mentionnés dans le paragraphe précédent, des réalisations de X selon la loi donnée par (3.1).

Notons que cette procédure, appelée « échantillonneur de Gibbs », est un cas particulier des démarches générales connues sous l'appellation « Méthodes de Monte Carlo par Chaînes de Markov » (MCMC), dont la problématique est la

suiuante. On souhaite simuler une loi $p(x)$ trop compliquée pour que l'on puisse le faire directement. On construit alors des probabilités de transitions $p(x^{i+1} | x^i)$, suffisamment simples pour permettre les simulations, définissant une chaîne de Markov dont les lois marginales tendent vers $p(x)$. Cette construction permet de faire des simulations de $p(x)$ « à la limite », en faisant un nombre suffisant de simulations selon les transitions.

La loi $p(x)$ étant définie par (3.1), il reste à définir les lois $p(y|x)$. Leur forme la plus simple, couramment utilisée, est

$$p(y|x) = \prod_{s \in S} f_{x_s}(y_s) = \exp \left[\sum_{s \in S} \text{Log}(f_{x_s}(y_s)) \right] \quad (3.3)$$

La loi de X a posteriori $p(x|y)$, qui est proportionnelle, à une constante dépendant de y près, à $p(x, y) = p(x)p(y|x)$, s'écrit alors

$$p(x|y) = \gamma(y) \exp[-U(x, y)] = \gamma(y) \exp \left[- \sum_{c \in C} \varphi_c(x_c) + \sum_{s \in S} \text{Log}(f_{x_s}(y_s)) \right] \quad (3.4)$$

Nous constatons que (3.4) diffère de (3.1) par la présence d'une somme sur les pixels ; les singletons étant toujours des cliques, la structure Markovienne donnée par (3.1) est donc préservée. Il est donc possible de simuler des réalisations de X selon sa loi a posteriori (3.4) en utilisant les lois conditionnelles (3.2), avec $\varphi_c(x_s, x_{\bar{c}}) - \text{Log}(f_{x_s}(y_s))$ à la place de $\varphi_c(x_s, x_{\bar{c}})$, et l'échantillonneur de Gibbs. En désignant par x^1, \dots, x^n les images simulées, les marginales a posteriori peuvent être aisément estimées par

$$\hat{p}(x_s = \omega | y) = \frac{1_{[x_s^1 = \omega]} + \dots + 1_{[x_s^n = \omega]}}{n} \quad (3.5)$$

et on dispose alors des outils nécessaires pour la mise en place des stratégies Bayésiennes de l'estimation de X à partir de Y . A titre d'exemple, la stratégie du Maximum des Marginales a Posteriori (MPM), qui est la décision Bayésienne d_1 correspondant à la fonction de perte $L_1(x^1, x^2) = \sum_{s \in S} 1_{[x_s^1 \neq x_s^2]}$, s'écrit :

$$\hat{x} = d_1(y), \text{ avec } \hat{x}_s = \arg \max_{\omega} p(x_s = \omega | y) \quad (3.6)$$

Nous voyons que (3.6) répond à la question posée et permet d'effectuer la recherche de chaque \hat{x}_s à partir de toute l'information disponible $Y = y$. Le calcul explicite des marginales a posteriori $p(x_s|y_s)$ (voir (2.4) de l'exemple 2.2) a été remplacé par une estimation, par une technique de type MCMC, des $p(x_s|y)$. Notons que la stratégie Bayésienne d_2 , correspondant à la fonction de perte $L_2(x^1, x^2) = 1_{[x^1 \neq x^2]}$, est donnée par

$$\hat{x} = d_2(y) = \arg \max_x p(x|y) \quad (3.7)$$

et peut être approchée soit par le très élégant algorithme de « recuit simulé » [16], soit par des algorithmes rapides comme l'algorithme « Iterated Conditional Mode (ICM [6]). Dans le premier, on introduit dans l'énergie U de (3.4) la température $T > 0$ en posant $U_T(x, y) = U(x, y)/T$. En choisissant une suite (T_n) décroissante et tendant vers 0, notons x_n la solution (3.7) associée à (3.4), avec U_{T_n} . On montre alors que si la décroissance de (T_n) est suffisamment lente, la suite (x_n) tend vers \hat{x} donné par (3.7). Dans les cas des modèles complexes, l'algorithme du recuit simulé peut être éventuellement accéléré par des techniques de parallélisation [24]. L'algorithme ICM ressemble à l'échantillonneur de Gibbs: on effectue des balayages de l'ensemble des pixels et pour chaque $s \in S$ on remplace la valeur courante x_s par une nouvelle valeur x_s' (avec, éventuellement, $x_s = x_s'$) obtenue par la maximisation de la probabilité $p(x_s|x_{V_s}, y)$ (dans l'échantillonneur de Gibbs, la nouvelle valeur est obtenue par le tirage aléatoire selon cette même probabilité). On obtient ainsi une suite déterministe x^1, \dots, x^n telle que la suite $p(x^1|y), \dots, p(x^n|y)$ est croissante; cependant, sa convergence vers \hat{x} donné par (3.7) n'est pas assurée.

Notons que la forme simple (3.3) des lois $p(y|x)$ est équivalente aux hypothèses

(H_1) $p(y_s|x) = p(y_s|x_s)$ et

(H_2) les variables (Y_s) sont indépendantes conditionnellement à X .

Ces hypothèses peuvent être affaiblies; en effet, si l'on remplace (H_1) par $p(y_s|x) = p(y_s|x_{c'})$, avec c' sous-ensemble de S de petite taille et contenant s , la Markovianité de la loi a posteriori (3.4),

suffisante pour utiliser l'échantillonneur de Gibbs qui permet la mise en place des décisions Bayésiennes, est préservée. Cependant, l'ensemble des cliques C est enrichi par l'ensemble C' de toutes les « cliques » c' et donc la Markovianité de X a posteriori est relative à un système de voisinages de taille plus importante que sa Markovianité a priori [19]. Notons également que des champs de Markov cachés plus perfectionnés permettent de traiter le problème de détection des bords dans les cas des classes texturées [17, 26].

3.3 Chaînes de Markov cachées

Supposons que S est ordonné et considérons, comme précédemment, un processus d'intérêt caché $X = (X_1, \dots, X_N)$ et un processus observé $Y = (Y_1, \dots, Y_N)$. On suppose que chaque X_i prend ses valeurs dans un ensemble fini de classes $\Omega = \{\omega_1, \dots, \omega_k\}$ et chaque Y_i prend ses valeurs dans R . Le processus X est une chaîne de Markov si pour tout $1 \leq n \leq N-1$, $p(x_{n+1}|x_1, \dots, x_n) = p(x_{n+1}|x_n)$. La loi de X est alors donnée par la loi de X_1 et les transitions $p(x_{n+1}|x_n)$. En posant les mêmes hypothèses (H_1) $p(y_i|x) = p(y_i|x_i)$ et (H_2) les variables (Y_i) sont indépendantes conditionnellement à X , nous avons

$$p(x, y) = p(x_1)p(y_1|x_1) \prod_{n=2}^N p(x_n|x_{n-1})p(y_n|x_n) \quad (3.8)$$

A la différence des champs de Markov cachés, où les différentes quantités d'intérêt doivent être estimées via des méthodes de type MCMC, dans le cas de chaînes de Markov cachées ces quantités peuvent être calculées. Plus précisément, posons $\alpha(x_i) = p(x_i, y_1, \dots, y_i)$, $\beta(x_i) = p(y_i, \dots, y_N|x_i)$. Ces quantités, appelées respectivement « probabilités forward » et « probabilités backward », sont calculables par les récursions suivantes [15, 41]:

$$\alpha(x_1) = p(x_1)p(y_1|x_1), \text{ et}$$

$$\alpha(x_{i+1}) = \sum_{x_i \in \Omega} \alpha(x_i)p(x_{i+1}|x_i)p(y_{i+1}|x_{i+1}), \text{ pour } 1 \leq i \leq N-1;$$

$$\beta(x_N) = 1, \text{ et}$$

$$\beta(x_i) = \sum_{x_{i+1} \in \Omega} \beta(x_{i+1})p(x_{i+1}|x_i)p(y_{i+1}|x_{i+1}), \text{ pour } 1 \leq i \leq N-1.$$

La loi $p(x|y)$ de X a posteriori est alors la loi d'une chaîne de Markov donnée par

$$p(x_1|y) = \alpha(x_1)\beta(x_1) / \sum_{x_1' \in \Omega} \alpha(x_1')\beta(x_1') ; \quad (3.9)$$

$$p(x_{i+1}|x_i, y) = \beta(x_{i+1})p(x_{i+1}|x_i)p(y_{i+1}|x_{i+1}) / \beta(x_i)$$

De plus, les marginales a posteriori sont données par

$$p(x_i|y) = \alpha(x_i)\beta(x_i) \sum_{x_i' \in \Omega} \alpha(x_i')\beta(x_i') \quad (3.10)$$

ce qui permet le calcul de la décision Bayésienne MPM. De plus, la solution de la décision Bayésienne du MAP est également calculable par l'algorithme de Viterbi [15]

En résumé, le modèle par chaînes de Markov cachées permet un calcul direct des quantités d'intérêt, alors que celui par champs de Markov cachés fait appel à des simulations itératives, souvent coûteuses en temps calcul, qui permettent d'estimer ces quantités.

3. 4 Arbres de Markov cachés

Soit S un ensemble d'indices et $X = (X_s)_{s \in S}$, $Y = (Y_s)_{s \in S}$, deux processus aléatoires, chaque X_s prenant ses valeurs dans un ensemble fini de classes $\Omega = \{\omega_1, \dots, \omega_k\}$ et chaque Y_s prenant ses valeurs dans R . Soit S^1, \dots, S^n une partition de S représentant différentes « générations ». A chaque $s \in S^i$ correspond $s^+ \subset S^{i+1}$ appelé l'ensemble des « enfants » de s , les ensembles des enfants des éléments de S^i formant une partition. Par ailleurs, on suppose que S^1 admet un seul élément appelé « racine », on note par s^{++} l'ensemble de tous les descendants de s , et par s^- son unique « père ». Le processus X est un arbre de Markov si sa loi est donnée par

$$p(x) = p(x_1) \prod_{s \in S - S^n} \prod_{t \in s^+} p(x_t|x_s) \quad (3.11)$$

En supposant une nouvelle fois les variables (Y_s) indépendantes conditionnellement à X et $p(y_s|x) = p(y_s|x_s)$, la loi $p(x, y)$ du couple (X, Y) s'écrit

$$p(x, y) = p(x_{s^1})p(y_{s^1}|x_{s^1}) \prod_{s \in S - S^n} \prod_{t \in s^+} p(x_t|x_s)p(y_t|x_t) \quad (3.12)$$

On montre alors que la loi de X a posteriori reste une loi d'arbre de Markov. De plus, les transitions « père-fils » a posteriori sont calculables ainsi que les solutions explicites des décisions Bayésiennes MPM et MAP [10, 27]

4. Modèles de Markov Couple

4.1 Textures

Considérons le cas d'une seule classe dans l'image présentant une « texture », que l'on peut modéliser, dans le cadre probabiliste considéré, comme la réalisation d'un champ aléatoire dont les composantes sont spatialement dépendantes. Si l'on suppose les champs aléatoires stationnaires au second ordre, il existe un certain nombre de modèles comme les modèles de Markov, CAR, AR, ou ARMA [19]. Les modèles de Markov sont parmi les modèles les plus simples. A titre d'exemple, supposons que l'image représente une forêt dont la texture est modélisée par un champ aléatoire $Y = (Y_s)_{s \in S}$ Gaussien et Markovien relativement aux quatre plus proches voisins. Sa loi s'écrit

$$p(y) = \frac{1}{\sqrt{\det(2\pi\Gamma)}} \exp\left[-\frac{y^T \Gamma^{-1} y}{2}\right] \quad (4.1)$$

$$= \frac{\exp\left[-\sum_{(t,u) \in C^1} \beta^1(y_t - y_u)^2 - \sum_{(t,u) \in C^{21}} \beta^2(y_t - y_u)^2\right]}{\sqrt{\det(2\pi\Gamma)}}$$

où C^1 est l'ensemble des pixels voisins horizontalement et C^2 est l'ensemble des pixels voisins verticalement.

4.2 Markov cachés et Markov couple

Considérons le problème de segmentation d'une image en deux classes texturées « eau » et « forêt ». En adoptant le modèle champs de Markov cachés, considérons X de Markov relativement aux quatre plus proches voisins et supposons que les deux textures sont modélisées par les lois de type (4.1). la loi de (X, Y) s'écrit

$$p(x, y) = p(x)p(y|x) = \frac{\gamma}{\sqrt{\det(2\pi\Gamma(x))}} \exp\left[-\sum_{c \in C^1 \cup C^2} \phi_c(x_c) - \sum_{(t,u) \in C^1} \beta_{x_t, x_u}^1(y_t - y_u)^2 - \sum_{(t,u) \in C^{21}} \beta_{x_t, x_u}^2(y_t - y_u)^2\right] = \gamma(x) \exp\left[-\sum_{c \in C^1 \cup C^2} \phi_c(x_c, y_c)\right] \quad (4.2)$$

On reconnaît l'écriture Markovienne sous l'exponentielle ; cependant, $\gamma(x) = \frac{\gamma}{\sqrt{\det(2\pi\Gamma(x))}}$

n'admet pas nécessairement une écriture Markovienne. Cela implique que $p(x|y)$ n'est pas nécessairement de Markov, ce qui pose problème car les traitements utilisent la Markovianité de $p(x|y)$. Lorsque l'on utilise les champs de Markov cachés, on est alors amené à faire diverses approximations, qui souvent reviennent à considérer que $\gamma(x)$ dans (4.1) ne dépend pas de x [25, 32].

En effet, la loi « approchée » $p'(x|y)$ est alors Markovienne ce qui rend les divers traitements, mentionnés dans la section précédente, possibles. Considérer un champ de Markov Couple consiste à poser directement la Markovianité du couple (X, Y) [38]. En reprenant l'exemple précédent, on peut considérer directement

$$p(x, y) = \lambda \exp\left[-\sum_{c \in C^1 \cup C^2} \phi_c(x_c, y_c)\right] \quad (4.3)$$

avec les ϕ_c définies dans (4.2). Nous constatons que dans le modèle Markov couple la loi $p(x|y)$ (qui est égale à la loi approchée $p'(x|y)$ obtenue dans le modèle Markov caché) est Markovienne, ce qui rend les traitements possibles sans approximation du modèle. Notons que dans le modèle de Markov Couple le processus X n'est pas nécessairement de Markov ; en effet, en multipliant et en divisant (4.3) par $\gamma(x)$ et en intégrant par rapport à y on trouve $p(x) = \frac{\lambda}{\gamma(x)} \exp\left[-\sum_{c \in C^1 \cup C^2} \phi_c(x_c, y_c)\right]$. Ainsi, par rapport au modèle « hiérarchique » classique on perd la Markovianité de $p(x)$, qui n'est pas nécessaire aux traitements, et on gagne la Markovianité de $p(x|y)$, qui est indispensable.

De manière analogue, on peut introduire les chaînes et les arbres de Markov Couple [39].

5. Apprentissage des Modèles de Markov

5.1 Champs de Markov

Supposons que la loi de X donnée par (3.1) dépend d'un ensemble des paramètres α , qui apparaissent dans les fonctions potentiel qui deviennent φ_c^α , et le problème est son estimation à partir de X . La méthode générale du Maximum de Vraisemblance (MV), dont le principe est

$$\hat{\alpha}(x) = \arg \max_{\alpha} \gamma(\alpha) \exp\left[-\sum_{c \in C} \varphi_c^\alpha(x_c)\right] \quad (5.1)$$

ne peut être appliquée directement car $\gamma(\alpha)$ est inconnue. La première idée est de remplacer la vraisemblance par la « pseudo-vraisemblance », qui est le logarithme du produit par rapport à $s \in S$ des lois conditionnelles (3.2). La fonction $\gamma(\alpha)$ disparaît et on peut, sous certaines conditions, calculer ou approcher le Maximum de la Pseudo-Vraisemblance (MPV). De plus, MPV jouit de bonnes propriétés asymptotiques. Une deuxième méthode, que nous appellerons « Estimateur Empirique » (EE) et qui peut être appliquée lorsque la taille du voisinage et le nombre de classes ne sont pas trop importants, consiste à calculer la fréquence d'apparition de chacune des classes conditionnellement à toutes les configurations du voisinage, et à rechercher α qui ajuste ces fréquences aux lois conditionnelles données par (3.2). Lorsque l'énergie dépend linéairement du vecteur $\alpha = (\alpha_1, \dots, \alpha_m)$, à savoir

$$U_\alpha(x) = \sum_{1 \leq i \leq m} \alpha_i \left[\sum_{c \in C_i} \varphi_c(x_c) \right] = \sum_{1 \leq i \leq m} \alpha_i U_i(x), \quad \text{cet}$$

ajustement peut être fait au sens des moindres carrés [9, 35].

Enfin, toujours sous l'hypothèse de la dépendance linéaire, on peut montrer que

$$\frac{\partial p_\alpha(x)}{\partial \alpha_i} = U_i(x) - E_\alpha[U_i(X)], \quad (5.2)$$

l'espérance $E_\alpha[U_i(X)]$ pouvant être approchée par la moyenne empirique calculée à partir des simulées de X . Il est alors possible de mettre en place des méthodes dites du gradient stochastique (GS) qui permettent de trouver, sous certaines conditions, l'estimateur du MV [48].

5.2 Champs de Markov cachés

Notons α l'ensemble des paramètres donnant la loi $p_\alpha(x)$ de X , β l'ensemble des paramètres donnant les lois conditionnelles $p_\beta(y|x)$, et $\theta = (\alpha, \beta)$ l'ensemble des paramètres que l'on cherche à estimer à partir de Y . Nous décrivons brièvement deux méthodes générales d'estimation dans le cas des données incomplètes.

La vraisemblance de la loi de Y , qui s'écrit $p_\theta(y) = \sum_{x \in \Omega^N} p_\alpha(x) p_\beta(y|x)$, est trop complexe

pour que l'on puisse envisager le calcul direct de l'estimateur MV. Le principe de la méthode « Expectation-Maximization » (EM [13, 42]), qui

permet de définir une suite (θ^n) telle que la suite $(p_{\theta^n}(y))$ est croissante, est le suivant :

- (i) on se donne un paramètre initial θ^0 ;
- (ii) θ^{q+1} est défini à partir de θ^q et y par

$$\theta^{q+1} = \arg \max_{\theta} E_{\theta^q} [\text{Log}(p_{\theta}(X, Y) | Y = y)] \quad (5.3)$$

Notons que l'appellation EM vient du fait que (5.3) donne généralement θ^{q+1} en deux temps : calcul de l'espérance E_{θ^q} (en anglais « expectation »), et sa maximisation. La méthode EM est très largement utilisée et donne généralement des résultats satisfaisants ; cependant, dans le contexte des champs de Markov cachés son application est malaisée et l'on doit s'écarter du principe général en utilisant des simulations stochastiques [9, 30, 31, 35, 40, 50].

La deuxième méthode générale, appelée « Iterative Conditional Estimation » (ICE [36]) est applicable dès que l'on dispose d'un estimateur de θ à partir des données complètes $\hat{\theta} = \hat{\theta}(X, Y)$ et d'une méthode de simulation de X selon $p_{\theta}(x|y)$. Le déroulement de ICE est le suivant :

- (i) on se donne un paramètre initial θ^0 ;
- (ii) θ^{q+1} est défini à partir de θ^q et y par

$$\theta^{q+1} = E_{\theta^q} [\hat{\theta}(X, Y) | Y = y] \quad (5.4)$$

lorsque cette espérance est calculable, et par

$$\theta^{q+1} = \frac{\hat{\theta}(x^1, y) + \dots + \hat{\theta}(x^k, y)}{k} \quad (5.5)$$

avec x^1, \dots, x^k simulés selon $p_{\theta^q}(x|y)$, lorsqu'elle ne l'est pas.

Remarquons qu'il peut arriver d'appliquer (5.4) pour certaines composantes de θ et (5.5) pour les composantes restant.

Nous avons vu que X pouvait être simulé selon $p_{\theta}(x|y)$ par l'échantillonneur de Gibbs, l'existence de $\hat{\theta} = \hat{\theta}(X, Y)$ est ainsi la seule condition à vérifier. C'est une condition très faible car, si l'on ne trouve pas un estimateur à partir des données complètes il est illusoire d'en chercher un à partir des seules données observées. Ainsi si $p_{\beta}(y|x)$ est donnée par (3.3) et si les k densités f_1, \dots, f_k sont paramétrées, on peut poser $\hat{\theta} = \hat{\theta}(X, Y) = (\hat{\alpha}(X), \hat{\beta}(X, Y))$, avec pour $\hat{\alpha}(X)$ l'un des estimateurs du paragraphe 5.1 et pour

$\hat{\beta}(X, Y)$ les estimateurs des moments empiriques, valables dans les cas des densités classiques sur R . ICE ne fait ainsi pas appel à la notion de vraisemblance et peut utiliser tout estimateur $\hat{\theta} = \hat{\theta}(X, Y)$, ce qui la rend relativement souple et générale. De par cette souplesse, ICE semble bien adaptée aux champs de Markov cachés et a été utilisée avec succès dans de nombreuses situations [7, 33, 34, 44].

Signalons également d'autres méthodes d'estimation, comme celle dérivant du gradient stochastique [49], ou celle de Lakshmanan et al. [28]. Signalons la possibilité de l'estimation des modèles plus généraux, où la nature des densités f_1, \dots, f_k n'est pas connue ; cependant, chacune appartient à un ensemble connu de types admissibles. Par exemple, chacune est Gaussienne ou Gamma, mais on ne sait pas dans quel cas, parmi les 2^k possibilités, l'on se trouve. On peut alors proposer des méthodes d'estimation généralisant EM et ICE, qui estiment α , donnent le type pour chacune des f_1, \dots, f_k , et estiment les paramètres β_i fixant f_i dans le type donné. De telles méthodes sont appelés estimateurs de « mélanges généralisés » [12].

5.3 Champs de Markov Couple

Notons brièvement la possibilité de l'utilisation de ICE dans le contexte des champs de Markov Couple. Le processus $Z = (X, Y)$ est de Markov, sa loi $p_{\theta}(x, y)$ dépend des paramètres θ que l'on cherche à estimer à partir de $Y = y$. Nous devons d'abord considérer un estimateur à partir des données complètes $\hat{\theta} = \hat{\theta}(X, Y) = \hat{\theta}(Z)$, ce qui est possible en prenant l'un des estimateurs MPV, EE, ou GS brièvement décrits dans le paragraphe 5.1 (notons bien que l'on considère ici $Z = (X, Y)$ à la place de X considéré dans le paragraphe 5.1). Par ailleurs, X étant Markovien conditionnellement à $Y = y$, ses simulations selon $p_{\theta}(x|y)$ sont possibles. Les deux conditions de l'ICE sont ainsi remplies.

5.4 Chaînes et arbres de Markov cachés et couple.

Dans les cas classiques des chaînes ou arbres de Markov cachés avec du bruit Gaussien, l'estimation des paramètres ne pose pas de problème majeur [5, 27]. A titre d'exemple, les formules de réestimation (5.3) définissant l'algorithme EM peuvent être explicitées analytiquement, ce qui rend l'estimation rapide. Par ailleurs, diverses études montrent que EM et ICE sont, lorsqu'il s'agit d'utiliser les paramètres estimés à des fins de classification Bayésienne, extrêmement performants. En d'autres termes, la qualité des segmentations effectuées sur

la base des paramètres estimés à partir de EM ou ICE est proche de celle des segmentations effectuées à partir des vrais paramètres, cela jusqu'à un niveau élevé du bruit. De plus, EM et ICE sont très robustes dans le sens où, lorsque les données ne suivent pas nécessairement un modèle de Markov caché, la qualité des segmentations effectuées sur la base des paramètres estimés à partir des seules données observées par EM ou ICE est proche de celle des segmentations effectuées sur la base des paramètres estimés à partir des données complètes [5]. Notons que EM présente un avantage par rapport à ICE au niveau du temps de calcul : dans itérations, les paramètres α de la loi de X sont réestimés par les mêmes formules analytiques mais les paramètres du bruit, réestimés par des formules analytiques dans EM, le sont à partir des simulations dans ICE.

Ces qualités des chaînes et arbres de Markov cachés sont remarquables et montrent qu'il reste une certaine marge permettant leur complexification. En effet, leurs généralisation aux modèles Couple est possible et les premières simulations, qui sont en cours, donnent des résultats encourageants.

6. Extensions

La robustesse des modèles de Markov cachés, aussi bien en ce qui concerne la robustesse des traitements par rapport à l'adéquation des modèles aux données qu'en ce qui concerne les traitements non supervisés par rapport aux traitements supervisés, permet de proposer leur utilisation dans des situations plus complexes, dont nous présentons brièvement ci-après une liste non exhaustive.

6.1 Images multisenseurs.

Dans le cas de plusieurs senseurs chaque Y_s prend ses valeurs dans R^m , situation à laquelle les différents modèles de Markov cachés se généralisent aisément lorsque les paramètres sont connus. Lorsque les bruits sont Gaussiens, ou lorsqu'ils ne sont pas Gaussiens mais les senseurs sont indépendants, l'estimation des paramètres ne pose pas de problème particulier, même lorsque la nature exacte des composantes du mélange n'est pas connue [18]. Lorsque les bruits sont non nécessairement Gaussiens et les senseurs sont dépendants, l'estimation est plus compliquée, notamment dans le cas où les types de densités ne sont pas connues [37].

6.2 Images 3D et séquences d'images.

Lorsque l'ensemble des pixels se trouve dans R^3 , on les appelle alors « voxels », on a une image 3D, ou une séquence d'images. Les champs de Markov cachés se généralisent aisément à cette situation [20, 24]; cependant, le nombre de paramètres croît rapidement et la durée des divers traitements utilisant des techniques MCMC peut devenir

réduisant. On peut alors imaginer d'utiliser les chaînes de Markov, après avoir converti l'ensemble des voxels en une suite, par des parcours de type Hilbert-Peano [18, 46]. L'introduction des modèles Couple est possible, de même que les traitements des images 3D – ou des séquences d'images – multisenseurs. Enfin, l'estimation des paramètres, dont le nombre cependant croît rapidement, ne pose pas de problèmes autres que ceux de la durée des calculs.

6.3 Images multirésolution

Reprenons la situation du paragraphe 3.4, où S^1, \dots, S^n est une partition de S représentant diverses « générations ». Il existe alors une grande diversité des cas possibles. Les variables X_s peuvent avoir une existence réelle pour certaines générations et fictive pour d'autres, leurs ensembles d'arrivée pouvant par ailleurs varier avec la génération. La Markovianité du processus de $X = (X_s)_{s \in S}$ peut alors être de type « arbre », avec des traitements utilisant des calculs directs et s'apparentant aux calculs classiques dans les chaînes de Markov [27, 43], ou de type « champ », où les calculs directs ne sont pas possibles et l'on doit faire appel à des techniques de simulation, généralement de type MCMC. Notons que le panachage de deux types de modèles est possible ; à titre d'exemple, on peut imaginer une pyramide où la restriction de X à S^1 , notée X^1 , qui est de forme rectangulaire, est un champ de Markov et la loi de sa restriction à $S - S^1$ est un arbre de Markov [21]. La loi de X est alors donnée par (3.12), avec la loi d'un champ Markovien $p(x^1)$ à la place de $p(x_1)$. En adoptant les hypothèses classiques de l'indépendance des (Y_s) conditionnellement à X et $p(y_s|x) = p(y_s|x_s)$, $p(x, y)$ garde la structure de $p(x)$, et il en est de même de $p(x|y)$. Cette dernière propriété rend possible soit le calcul direct de diverses quantités d'intérêt, soit leurs estimation à partir des simulées de X selon $p(x|y)$ via des méthodes MCMC. Les quantités d'intérêt en question permettent alors aussi bien l'estimation des paramètres par des méthodes comme EM ou ICE que les segmentations Bayésiennes correspondant à différentes fonctions de perte. Enfin, la généralisation aux modèles Couple, où on attribuerait à $Z = (X, Y)$ la structure Markovienne de X , est envisageable.

6.4 Fusion de Dempster-Shafer

Mentionnons brièvement une possibilité de généralisation des modèles Markoviens. En se plaçant sur un pixel et reprenant l'exemple d'eau et forêt de la section 2, nous avons vu que la probabilité

a posteriori, qui permet la segmentation Bayésienne, est

$$p(E|y_s) \propto p(E)p(y_s|E),$$

$$p(F|y_s) \propto p(F)p(y_s|F).$$

En considérant la probabilité

$$q(E) \propto p(y_s|E), \quad q(F) \propto p(y_s|F),$$

nous constatons que la probabilité a posteriori est obtenue à partir de la probabilité a priori $p(E)$, $p(F)$ et la probabilité q donnée par l'observation par la « fusion »

$$p(E|y_s) \propto p(E)q(E), \quad p(F|y_s) \propto p(F)q(F).$$

Il s'agit là d'un cas particulier de la fusion de Dempster-Shafer (DS), qui s'applique dans un cadre plus général [1, 4, 45, 47]. Supposons que nous sommes en présence de la classe « nuages » notée N , avec $p(y_s|N)$, ce qui donne une probabilité q

définie sur $\{E, F, N\}$ (ou encore sur $\{E, F, \{E, F\}\}$ car observer N ne nous donne aucune information sur les classes d'intérêt E, F ; dans le langage de la fusion DS q est dite « fonction de masses ») par

$$q(E) \propto p(y_s|E), \quad q(F) \propto p(y_s|F), \quad \text{et}$$

$$q(N) \propto p(y_s|N).$$

La fusion DS donne alors

$$p(E|y_s) \propto p(E)[q(E) + q(N)],$$

$$p(F|y_s) \propto p(F)[q(F) + q(N)],$$

qui est une probabilité « généralisant » la probabilité a posteriori et qui peut être utilisée à des fins de classification.

Cette démarche se généralise au cas de m senseurs indépendants, certains d'entre eux pouvant être sensibles à d'autres classes que celles d'intérêt, d'autres pouvant être insensibles à certaines classes d'intérêt ... Ce type de modélisation peut être appliquée dans les modèles de Markov, les premières études concernant les champs [4] ou chaînes [14] de Markov cachés.

6.5 Données incomplètes

Dans ce qui précède nous avons supposé tous les (Y_s) observés et tous les (X_s) recherchés, situation qui porte parfois le nom du problème des « données cachées ». La situation des « données incomplètes » plus générale, peut encore être traitée en utilisant les modèles de Markov. Dans une telle situation, supposons que X est observé sur $S_1 \subset S$ et Y est observé sur $S_2 \subset S$ (on retrouve la situation précédente avec $S_1 = \emptyset$ et $S_2 = S$). On peut alors mettre en place, dès que l'on dispose d'un modèle de Markov caché pour la loi de (X, Y) , des méthodes Bayésiennes d'estimation de X_{S-S_1} à

partir de Y_{S_2} . Notons que cela est en particulier vrai pour $S_1 = \emptyset$ et $S_2 \neq S$: on peut ainsi estimer les réalisations des X_s même pour les pixels éloignés des éléments de S_2 sur lequel on observe les Y_t (rappelons que dans un champ de Markov $X = (X_s)_{s \in S}$ les variables X_t, X_u sont toujours dépendantes, quelles que soient les positions de t et u dans S).

7. Conclusions et Perspectives

Les traitements statistiques d'images fondés sur des modèles Markoviens peuvent présenter des qualités exceptionnelles. L'avantage de ces modèles par rapport à des modèles « locaux » découle de leur aptitude à prendre en compte, de façon souvent élégante et mathématiquement rigoureuse, l'ensemble de l'information disponible. De plus, les diverses études semblent indiquer qu'une extraordinaire robustesse s'ajoute aux qualités classiques des méthodes statistiques que sont la souplesse et l'optimalité. Cette robustesse permet d'envisager des complexifications croissantes des modèles : séquences d'images, images 3D, multirésolutions, multisenseurs avec senseurs corrélés et non nécessairement Gaussiens, utilisation de la théorie de l'évidence et du flou,

Notons que les champs, chaînes, ou arbres de Markov présentés dans cet article font parties des « modèles graphiques » plus généraux, faisant actuellement l'objet d'une activité de recherche importante [3, 22, 23]. Les divers résultats obtenus en dehors des préoccupations liées aux traitements d'images seront sans doute parmi les outils les plus stimulants dans la recherche des diverses généralisations des modèles utilisés en imagerie.

Références

- [1] A. Appriou, Probabilités et incertitude en fusion de données multisenseurs, *Revue Scientifique et Technique de la Défense*, No. 11, pp. 27-40, 1991.
- [2] R. Azencott, Simulated Annealing : Parallelization Techniques, Wiley, 1992.
- [3] A. Becker et P. Naïm, Les Réseaux Bayésiens, Eyrolles, Paris, 1999.
- [4] A. Bendjebbour, Y. Delignon, L. Fouque, V. Samson, and W. Pieczynski, Multisensor Images Segmentation Using Dempster-Shafer Fusion in Markov Fields Context, *IEEE Transactions on GRS*, Vol. 39, No. 8, pp. 1789-1798, 2001.
- [5] B. Benmiloud, W. Pieczynski, Estimation des paramètres dans les chaînes de Markov cachées et segmentation d'images, *Traitement du Signal*, Vol. 12, No. 5, pp. 433-454 1995.

- [6] J. Besag, On the statistical analysis of dirty pictures, *JRSS, Series B*, 48, pp. 259-302, 1986.
- [7] J. M. Boucher and P. Lena, Unsupervised Bayesian classification, application to the forest of Paimpont (Brittany), *Photo Interprétation*, Vol. 32, No. 1994/4, 1995/1-2, pp. 79-81, 1995.
- [8] E. Bratsolis and M. Sigelle. Image relaxation using Potts model with a fast deterministic method. *JOSA, A*, pp. 1033-1043, 1997.
- [9] B. Chalmond, *Eléments de modélisation pour l'analyse d'images*, Springer, SMAI, 2000.
- [10] C. Collet, J.-N. Provost, P. Rostaing, P. Pérez, P. Bouthemy. Segmentation bathymétrique d' images multispectrales SPOT, *Traitement du Signal*, 18(1), pp. 1-16, 2001.
- [11] Descombes X., Morris R. D., Zerubia J., Berthod M., "Estimation of Markov Random Field Parameters Using Markov Chain Monte Carlo Maximum Likelihood", *IEEE Transactions on IP*, Vol. 8, No. 7, pp. 954-963, 1999.
- [12] Y. Delignon, A. Marzouki and W. Pieczynski, "Estimation of generalised mixtures and its application in image segmentation", *IEEE Transactions on IP*, Vol. 6, No. 10, pp. 1364-1375, 1997.
- [13] J.-P. Delmas, An equivalence of the EM and ICE algorithm for exponential family, *IEEE Transactions on SP*, Vol. 45, No. 10, pp. 2613-2615, 1997.
- [13] A. P. Dempster, N.M. Laird, and D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, *JRSS B*, 39, pp. 1-38, 1977.
- [14] L. Fouque, A. Appriou, and W. Pieczynski, An Evidential Markovian Model for Data Fusion and Unsupervised Image Classification, *Proceedings of 3rd International Conference on Information Fusion, FUSION 2000*, Vol. 1, July 10th-13th, 2000, Paris, France, pp. TuB4-25 - TuB4-31.
- [15] G.D. Fornay, The Viterbi algorithm, *Proceedings of the IEEE*, Vol. 61, No. 3, pp. 268-277, 1973.
- [16] S. Geman and D. Geman, Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images, *IEEE Transactions on PAMI*, Vol. 6, pp. 721-741, 1984.
- [17] D. Geman, S. Geman, C. Graffigne, and P. Dong, Boundary detection by constrained, *IEEE Transactions on PAMI*, Vol. 12 No: 7, pp. 609 -628, 1990.
- [18] N. Giordana and W. Pieczynski, Estimation of Generalized Multisensor Hidden Markov Chains and Unsupervised Image Segmentation, *IEEE Transactions on PAMI*, Vol. 19, No. 5, pp. 465-475, 1997.
- [19] X. Guyon, *Random Fields on Network. Modeling, Statistics, and Applications*. Springer-Verlag, Probability and its Applications, New York, 1995.
- [20] F. Heitz, P. Bouthemy. Multimodal estimation of discontinuous optical flow using Markov random fields. , *IEEE Transactions on PAMI*, Vol. 15, No. 12, pp. 1217-1232, 1993.
- [21] L. Hubert-Moy, A. Cotonnec, L. Le Du, A. Chardin, P. Pérez. A comparison of parametric classification procedures of remotely sensed data applied on different landscape units. *Remote Sensing Environment*, 75(2):174-187, 2001.
- [22] F. V. Jensen, *An Introduction to Bayesian Networks*, UCL Press, 2000.
- [23] M. I. Jordan, *Learning in Graphical Models*, Kluwer Academic Publishers, 1998.
- [24] Z. Kato, J. Zeroubia, and M. Berthod, Unsupervised parallel image classification using Markovian models, *Pattern Recognition*, Vol. 32, pp. 591-604, 1999.
- [25] P.A. Kelly, H. Derin, and K.D. Hartt, Adaptive segmentation of speckled images using a hierarchical random field model, *IEEE Transactions on ASSP*, Vol. 36, No. 10, pp. 1628-1641, 1988.
- [26] C. Kervrann, F. Heitz. A Markov random field model-based approach to unsupervised texture segmentation using local and global spatial statistics. *IEEE Transactions on IP*, Vol. 4, No. 6, pp. 856-862, 1995.
- [27] J.-M. Laferté, P. Pérez, and F. Heitz, Discrete Markov image modeling and inference on the quadtree, *IEEE Transactions on IP*, Vol. 9 , No 3, pp. 390 - 404, 2000.
- [28] S. Lakshmanan, and H. Derin, Simultaneous parameter estimation and segmentation of Gibbs random fields, *IEEE Transactions on PAMI*, Vol. 11, pp. 799-813, 1989.
- [29] J. Marroquin, S. Mitter, T. Poggio, Probabilistic solution of ill-posed problems in computational vision, *JASA*, 82, pp. 76-89, 1987.
- [30] G. J. McLachlan, *Discriminant Analysis and Statistical Pattern Recognition*, Wiley, 1992.
- [31] G. J. McLachlan and T. Krishnan, *EM Algorithm and Extensions*, Wiley, 1997.
- [32] D. Melas and S. P. Wilson, *Double Markov Random Fields and Bayesian Image Segmentation*,

IEEE Transactions on SP, Vol. 50, No. 2, pp. 357-365, 2002.

[33] M. Mignotte, C. Collet, P. Pérez, and P. Bouthémy, Sonar image segmentation using an unsupervised hierarchical MRF model, *IEEE Transactions on IP*, Vol. 9, No. 7, pp. 1216-1231, 2000.

[34] M. Mignotte, J. Meunier, J.-P. Soucy, C. Janicki. Comparison of deconvolution techniques using a distribution mixture parameter estimation: Application in single photon emission computed tomography imagery, *Journal of Electronic Imaging*, 11(1) , 2001.

[35] P. Pérez. Markov random fields and images. *CWI Quarterly*, 11(4):413-437, 1998.

[36] W. Pieczynski, Statistical image segmentation, *Machine Graphics and Vision*, Vol. 1, No. 1/2, pp. 261-268, 1992.

[37] W. Pieczynski, J. Bouvrais, and C. Michel, Estimation of Generalized Mixture in the Case of Correlated Sensors, *IEEE Transactions IP*, Vol. 9, No. 2, pp. 308-311, 2000.

[38] W. Pieczynski and A.-N. Tebbache, Pairwise Markov random Fields and segmentation of Textured Images, *Machine Graphics and Vision*, Vol. 9, No. 3, pp. 705-718, 2000.

[39] W. Pieczynski, Pairwise Markov Chains and Bayesian Unsupervised Fusion, Proceedings- of 3rd International Conference on Information Fusion, FUSION 2000, Vol. 1, July 10th-13th, 2000, Paris, France, pp. MoD4-24 - MoD4-31.

[40] W. Qian, and D. M. Titterton, Stochastic relaxations and EM algorithms for Markov random fields, *J. Statist. Comput. Simulation*, 40, pp. 55-69, 1992.

[41] L.R. Rabiner. A tutorial on hidden Markov models and selected applications in speech

recognition, *Proceedings of IEEE*, Vol. 77, No. 2, pp. 257-286, 1989.

[42] R.A. Redner and H.F. Walker, Mixture densities, maximum likelihood and the EM algorithm, *SIAM Review*, 26, pp. 195-239, 1984.

[43] P. Rostaing, J.-N. Provost, and C. Collet, Unsupervised Multispectral Image Segmentation using Generalized Gaussian Noise Model, Proceedings of EMMCVPR'99, York, England, July 1999.

[44] F. Salzenstein and W. Pieczynski, Parameter Estimation in Hidden Fuzzy Markov Random Fields and Image Segmentation, *GMIP*, Vol. 59, No. 4, pp. 205-220, 1997.

[45] G. Shafer, A mathematical theory of evidence, Princeton University Press, Princeton, 1976.

[46] W. Skarbek, "Generalized Hilbert scan in image printing", *Theoretical Foundations of Computer Vision*, R. Klette and W. G. Kropetsh, editors, pp.45-57, Akademik Verlag, 1992.

[47] P. Smets, Belief functions: the disjunctive rule of combination and the generalized Bayesian theorem, *International Journal of Approximate Reasoning*, Vol. 9, pp. 1-35, 1993.

[48] L. Younes, "Estimation and annealing for gibbsian fields," *A. Inst Henri Poincaré*, 24(2):269—294, 1988.

[49] L. Younes, "Parametric inference for imperfectly observed gibbsian fields," *Probability Theory and Related Fields*, 82:625—645, 1989.

[50] J. Zhang, J. W. Modestino, and D. A. Langan, Maximum likelihood parameter estimation for unsupervised stochastic model-based image segmentation, *IEEE Transactions on IP*, Vol. 3, No. 4, 1994, pp. 404-420.

