

Modélisation de points imprimés à l'échelle micro-métrique

Quoc Thong NGUYEN¹, Yves DELIGNON¹, Lionel CHAGAS², François SEPTIER¹

¹Institut Mines Telecom/Telecom Lille/ CRISTAL UMR CNRS 9189
rue Guglielmo Marconi, BP 20145, 59653 Villeneuve d'Ascq, France

²Laboratoire Traitement des Images
461 rue de la Papeterie, CS 10065, 38402 Saint Martin d'Hères Cedex, France
quoc-thong.nguyen@telecom-lille.fr, delignon@telecom-lille.fr,
lionel.chagas@pagora.grenoble-inp.fr
septier@telecom-lille.fr

Résumé – La reconnaissance de caractères, le halftoning ou encore l'authentification de documents requièrent la modélisation des dégradations de l'impression à une échelle appropriée. Dans cet article, nous sommes focalisés sur la modélisation d'impression à l'échelle micrométrique et sur son estimation. L'image est considérée comme un ensemble de variables aléatoires de Bernoulli paramétrées par un noyau qui tient compte de la variabilité de forme des points imprimés. Ses caractéristiques dépendent du type d'imprimante, de son réglage, de la qualité de l'encre et des propriétés du papier. Nous développons un algorithme de Metropolis Hasting within Gibbs comme estimateur Bayésien du modèle. La borne de Cramer Rao a posteriori est aussi calculée, elle sert de référence à l'évaluation des performances de l'estimateur. Enfin nous comparons les performances du modèle et de son estimateur sur des impressions issues d'imprimantes laser, jet d'encre, offset waterless et offset conventionnel.

Abstract – Optical character recognition, halftoning or document authentication require the modelling of the printing degradations. This paper is focused on both the modelling of printings at the micrometer scale and on the model estimation. We propose to model a print as a set of independent Bernoulli variables parametrized by a kernel which takes into account the shape variability of printed dots. The kernel characteristics depend on the printer technology, its setup, the quality of the ink and the properties of the paper. We develop a Metropolis Hasting within Gibbs algorithm as Bayesian estimator of the model. The posterior Cramer Rao Bound is also calculated, it enables to assess the performances of the estimator. At end, the adequacy of both the model and its estimator is studied on prints obtained by laser, inkjet, conventional offset and waterless offset printers.

1 Introduction

Le contrôle de la qualité de l'impression est une problématique que l'on rencontre dans de nombreuses applications telles que le halftoning, l'authentification ou encore la reconnaissance de caractères. Plusieurs modèles de dégradations d'impression papier ont fait l'objet de publications scientifiques [1, 2]. Chaque modèle a été construit en fonction d'une application particulière à une échelle donnée.

A l'échelle micrométrique, les points imprimés, aussi appelés dot, ont une forme très variable dépendant de la technologie et du réglage des imprimantes, du support d'impression et des propriétés de l'encre (cf. Tab. 4). Cette signature présente un intérêt pour améliorer les performances de méthodes d'authentification de contrefaçons tels que des faux papiers [3], [4].

Dans cet article, nous proposons de modéliser l'impression scannée par un champ aléatoire constitué de dots aléatoirement répartis dans l'image. La variabilité de forme des dots est prise en compte par un noyau gaussien généra-

lisé. Les paramètres du modèle représentent alors la signature de l'impression, leur estimation est une tâche délicate qui a été abordée sous l'angle du maximum de vraisemblance dans [5]. Nous proposons ici un estimateur Bayésien basé sur l'algorithme de Metropolis Hasting within Gibbs (MHwG). Les performances de l'algorithme sont comparées à la borne de Cramer Rao a posteriori qui est aussi estimée. Enfin, le modèle et son estimateur sont évalués sur des impressions offset conventionnel, offset waterless, laser et jet d'encre.

2 Modèle binaire spatial

A l'échelle micrométrique, chaque point est un nuage de particules d'encre dont la forme et le nombre dépendent de la technologie de l'imprimante ainsi que de son réglage. Considérons l'image binaire $\mathbf{U} = (U_s)_{s \in S}$ comportant K points imprimés indépendants. \mathbf{U} est un champ aléatoire constitué de $N \times N$ pixels à valeur dans $\{0, 1\}^{N \times N}$ où 0 et 1 représentent respectivement les pixels noirs et blancs. Le paramètre de chaque variable aléatoire de Bernoulli U_s

dépend de la distance du pixel s au centre des points imprimés $\{\mu_k, k \in \{1, \dots, K\}\}$. L'étalement d'un dot k est fonction de $p_k(s) = \mathbb{P}(U_{s,k} = 0)$, $s \in S$ qui définit noyau du dot k dans l'espace image S . Puisque qu'un point imprimé k est un ensemble de particules d'encre agrégées autour de son centre, $p_k(s)$ décroît de manière continue avec la distance de s au centre μ_k .

Pour une image comportant K points imprimés, un pixel quelconque U_s est blanc si aucune particule d'encre n'impacte le site s parmi les K dots. Dans ce cas,

$$\mathbb{P}(U_s = 0) = 1 - \prod_{k=1}^K (1 - p_k(s)) \quad (1)$$

Supposant que le champ aléatoire est formé de pixels indépendants, la distribution de l'image U est donnée par :

$$q(\mathbf{u}) = \prod_s \left(1 - \prod_{k=1}^K (1 - p_k(s)) \right)^{\mathbb{I}_{S_0}(s)} \left(\prod_{k=1}^K (1 - p_k(s)) \right)^{\mathbb{I}_{S_1}(s)} \quad (2)$$

avec $S_i = \{s | u_s = i\}$ et $\mathbb{I}_A(\cdot)$ la fonction indicatrice sur A . Afin d'englober une grande variété de forme des points imprimés, nous proposons pour $p_k(\cdot)$ un noyau Gaussien généralisé :

$$p_k(s) = \eta \exp \left[-\frac{1}{2} ((s - \mu)' \Sigma^{-1} (s - \mu))^\beta \right], \quad (3)$$

où η est l'amplitude du noyau, $\eta \in (0, 1]$, β est le paramètre de forme du noyau, $\beta > 0$ conditionne l'applatissement noyau en son centre et Σ est une matrice symétrique définie positive. Dans la suite, on suppose Σ circulaire, soit $\Sigma = \sigma^2 I_2$ avec I_2 la matrice identité de dimension 2. η , Σ et β sont supposés identiques pour tous les points imprimés, c'est à dire pour tout $k \in \{1, \dots, K\}$. Remarquons que pour $\beta = 1$, (3) est le noyau Gaussien classique, et lorsque β croît, le noyau s'applatit en son sommet (cf. Fig. 1). Pour notre application, β joue un rôle important pour distinguer les formes des points imprimés à l'échelle micrométrique en fonction des imprimantes et de leur technologie. Sous forme vectorielle, l'ensemble des paramètres est noté $\Theta_{\mathbf{U}} = [\{\mu_k\}, \Sigma, \beta, \eta]$.

Une caractéristique importante de ce modèle est le nombre moyen de pixels noirs par point imprimé n :

$$n = \sum_s p_1(s) \simeq \eta \pi \sigma^2 2^{\frac{1}{\beta}} \Gamma \left(1 + \frac{1}{\beta} \right) \quad (4)$$

3 Estimation

Dans ce paragraphe, nous décrivons une méthode d'estimation Bayésienne des paramètres du modèle. Dans ce cas, le vecteur de paramètres $\Theta_{\mathbf{U}} = [\{\mu_k\}, \sigma^2, \beta, \eta]$ est aléatoire. Pour une observation u , l'estimation de $\Theta_{\mathbf{U}}$ est obtenue par la minimisation de l'erreur quadratique moyenne :

$$\hat{\theta}(\mathbf{U}) = \int \theta p(\theta | \mathbf{U}) d\theta \quad (5)$$

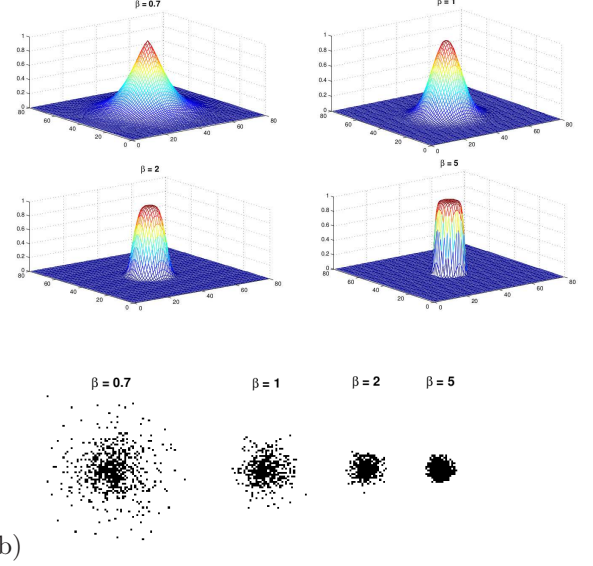


FIGURE 1 – a) $p_k(\cdot)$ avec $\sigma^2 = 50$, $\eta = 1$ et $\beta \in \{0.7; 1; 2; 5\}$ b) Réalisations du modèle binaire spatial pour $\sigma^2 = 50$, $\eta = 1$ et $\beta \in \{0.7; 1; 2; 5\}$.

où $p(\Theta | \mathbf{u})$ est la distribution a posteriori. L'intégrale étant incalculable analytiquement, on a recours à une méthode de Monte Carlo par Chaîne de Markov, l'algorithme de Metropolis Hasting within Gibbs. L'estimateur empirique est alors calculé sur la séquence simulée de longueur N_{iter} à laquelle est soustraite la période de chauffe $N_{chauffe}$. La méthode d'échantillonnage choisie est la marche aléatoire gaussienne. Le vecteur gaussien indépendant $\theta = [\theta_1, \theta_2, \theta_3, \theta_4]$ est formé de θ_1 qui est composé des centres des points $\{\mu_k\}$, $k \in \{1, \dots, K\}$, de $\theta_2 = \log(\sigma^2)$ and $\theta_3 = \log(\beta)$ puisque σ^2 et β sont positifs. η étant compris entre 0 et 1, la variable gaussienne associée est $\theta_4 = \log \left(\frac{1}{\eta} - 1 \right)$. L'algorithme de MHwG est décrit dans Algorithme 1 avec $\theta_{-b}^{(i-1)} = [\theta_1^{(i)}, \dots, \theta_{b-1}^{(i)}, \theta_{b+1}^{(i-1)}, \dots, \theta_4^{(i-1)}]$.

Bien connu comme sensible à l'initialisation des paramètres, l'agorithme MHwG est initialisé par l'algorithme Kmean ++ [6] pour les centres des dots μ_k , par $\frac{n}{K}$ pour le nombre moyen de pixels noirs par dot (n_b nombre de pixel noirs de l'image observée u), par $0.8 \times \frac{n^{(0)}}{\pi}$ pour $\sigma^{2(0)}$ et par l'équation (4) pour $\eta^{(0)}$.

Les distributions a priori des centres $\{\mu_k\}$ sont des variables uniformément réparties sur S . σ^2 étant positif, sa loi a priori est choisie lognormale, β est une variable gamma $\mathcal{G}(\phi, \alpha)$, alors que η est distribuée suivant une loi beta de première espèce de support $[0, 1]$, $\phi \sim \mathcal{B}(p, q)$.

Algorithm 1 Algorithme de Metropolis Hasting.

Initialisation $\theta^{(0)} = [\theta_1^{(0)}, \theta_2^{(0)}, \theta_3^{(0)}, \theta_4^{(0)}]$ avec $\theta_1^{(0)}$ obtenu par l'algorithme Kmeans++, $\theta_2^{(0)} = \log\left(0.8 \frac{n_b}{K\pi}\right)$, avec n_b nombre de pixels noirs dans u , $\theta_3^{(0)} = \log(\log(2)) - \log\left(\log\left(\frac{n_b}{K\pi}\right) - \theta_2^{(0)}\right)$, $\theta_4^{(0)} = \log\left(\frac{K\pi e^{\theta_2^{(0)}} 2^{\exp(-\theta_3^{(0)})} \Gamma(1+e^{-\theta_3^{(0)}})}{n_b} - 1\right)$

Pour $i = 1, \dots, N_{iter}$

Pour $b = 1, \dots, 4$

Générer $\theta_b^* \sim \mathcal{N}(\theta_b^{(i-1)}, \sigma_b^2)$

Calcul du rapport de probabilité

$$\alpha(\theta_b^{(i-1)}, \theta_b^*) = \min\left\{\frac{p(\theta_b^*|\theta_{-b}^{(i-1)}, \mathbf{u})}{p(\theta_b^{(i-1)}|\theta_{-b}^{(i-1)}, \mathbf{u})}, 1\right\}$$

Choisir

$$\theta_b^{(i)} = \begin{cases} \theta_b^* & \text{avec la probabilité } \alpha(\theta_b^{(i-1)}, \theta_b^*) \\ \theta_b^{(i-1)} & \text{autrement} \end{cases}$$

fin pour

fin pour

$$\hat{\theta} = \frac{1}{N_{iter} - N_{chauffe} + 1} \sum_{i=N_{chauffe}}^{N_{iter}} \theta^{(i)}$$

$$\mu_{1:K} = \hat{\theta}_1, \hat{\sigma}_2 = \exp \hat{\theta}_2, \hat{\beta} = \exp \hat{\theta}_3, \hat{\eta} = \frac{1}{1 + \exp \hat{\theta}_4}$$

4 Borne de Cramer Rao a posteriori

L'évaluation des performances des estimateurs est en général mesurée par leur biais et leur variance. Dans le cas d'un estimateur sans biais, les variances des estimateurs Bayésien sont bornées par les éléments diagonaux de l'inverse de la matrice d'information de Fisher a posteriori [7]

$$\text{var}(\Theta_i) \geq [J^{-1}]_{i,i} \quad (6)$$

avec la matrice d'information de Fisher :

$$J = -E_{X,\Theta} [\Delta_{\Theta}^{\Theta} \log P(X, \Theta)] \quad (7)$$

En réécrivant la loi du couple sous la forme du produit de la vraisemblance $P(X | \Theta)$ par la priori $f(\Theta)$, l'information de Fisher se décompose sous la forme :

$$J = -E_{\Theta} [J(\Theta) + \Delta_{\Theta}^{\Theta} \log f(\Theta)] \quad (8)$$

avec $J(\Theta) = E_{X|\Theta} [\Delta_{\Theta}^{\Theta} \log P(X | \Theta)]$ qui s'écrit après calculs :

$$J(\Theta) = \sum_{s \in S} \frac{\nabla_{\Theta} p_k(s) \nabla_{\Theta}^T p_k(s)}{p_k(s)(1 - p_k(s))} \quad (9)$$

quant à $\Delta_{\Theta}^{\Theta} \log f(\Theta)$, c'est une matrice diagonale, nulle pour les éléments diagonaux associés à $\{\mu_k\}$, valant

$$\frac{1}{\sigma^2} \left(1 + \frac{\log(\sigma^2) - \mu - 1}{\sigma^2}\right) \text{ pour } \sigma^2, \frac{p}{\eta^2} - \frac{q}{(1-\eta)^2} \text{ pour } \eta, \text{ et } \frac{1-\alpha}{\beta^2}$$

pour le paramètre de forme du noyau Gaussien généralisé β . L'espérance suivant Θ n'étant pas calculable analytiquement, l'information de Fisher J est obtenue par simulation de Monte carlo de l'équation (8). Un échantillon de

TABLE 1 – Borne de Cramer-Rao a posteriori (PCRB) et variance des estimateurs pour $\beta \in \{0.8, 2, 3, 5\}$.

β	$\hat{\sigma}^2$		$\hat{\beta}$		$\hat{\eta}$ (unit 10^{-3})	
	PCRB	Var.	PCRB	Var.	PCRB	Var.
0.8	0.74	116	0.001	0.02	1.8	6.6
2	0.71	28	0.025	0.23	1.6	6.4
3	0.65	4.5	0.104	0.11	0.7	1.2
5	0.55	1.6	0.569	0.69	0.2	0.8

$\Theta_{1:N}$ est simulé à partir des lois a priori et J est approchée par :

$$\hat{J} = -\frac{1}{N} \sum_{i=1}^N J(\Theta^{(i)}) + \Delta_{\Theta}^{\Theta} \log f(\Theta^{(i)}) \quad (10)$$

5 Performances

5.1 Etude de la précision de l'estimateur

Images d'un seul point imprimé

Les performances de l'algorithme MHwG sont analysées sur des images d'un seul point imprimé de forme variée. Un échantillon de 100 images de dimension 50×50 , avec $\mu = (25, 25)$, $\sigma^2 = 30$, $n = 95$ sont simulées pour $\beta \in \{0.8, 2, 3, 5\}$ représentatif de la variété de forme du noyau Gaussien généralisé. Obtenues par simulation de Monte Carlo, les variances des estimateurs de μ_1 , β et σ^2 sont comparées à leur borne de Cramer Rao a posteriori. L'algorithme est paramétré de la manière suivante, le nombre d'itération de l'algorithme est fixé à 10000 et la période de chauffe à 3000.

Tab. 1 montre que les estimateurs de σ^2 et η sont d'autant plus performants que β est grand contrairement à l'estimation de β . Lorsque β croît, l'information de Fisher a posteriori décroît, ce qui se traduit par une sensibilité décroissante de la forme du noyau aux variations de β pour des valeurs supérieures à 3 et en conséquence une imprécision de plus en plus forte de son estimation.

Images de K points imprimés

Les performances de l'algorithme sont étudiées ici sur des images comportant plusieurs dots. L'expérience est menée sur 100 images simulées comportant K dots séparés, $K \in \{1, 2, 4, 7\}$ avec $\sigma^2 = 50$, $\beta = 2$ et $n = 165$. Comme le montre le tableau Tab. 2, plus K est grand, plus les estimateurs sont performants.

5.2 Performances sur des données imprimées

L'expérience est réalisée sur des impressions issues des quatre technologies ; offset conventionnelle, offset water-

TABLE 2 – Variance des estimateurs de σ^2 , β et n pour $K \in \{1, 2, 4, 7\}$, $\sigma^2 = 50$, $\beta = 2$, $n = 165$ for each dot.

K	1	2	4	7
σ^2	24.8	11	4.6	6.5
β	0.044	0.036	0.02	0.006
n	59	32	16	9

TABLE 3 – Moyenne (Moy.) et écart-type (E. T.) de l’algorithme MHWG obtenus sur des impressions offset conventionnel, offset waterless, laser et jet d’encre.

	σ^2		β		n	
	Moy.	E. T.	Moy.	E. T.	Moy.	E. T.
C. offset	24.88	9.2	3.5	1.3	84	29
W. offset	126	40	4.85	3.84	469	115
Laser	84	74	1.17	0.56	324	105
Jet d’encre	705	138	3.36	0.77	2468	459

less, laser et jet d’encre. Les résolutions d’impression sont 1200 dpi pour les imprimantes offset, 600 dpi pour l’imprimante laser et 720 dpi pour l’imprimante jet d’encre. Pour chaque imprimante, 10 images d’un dot sont scannées. L’estimation par l’algorithme MHWG est réalisée avec 20 000 itérations MCMC, la séquence de chauffe est 5000.

Tab. 3 présente les performances de l’algorithme MHWG. Bienque les performances soient dégradées comparées aux données simulées à partir du modèle, les images simulées à partir des paramètres estimés sont visuellement proches des impressions (cf. Tab. 4).

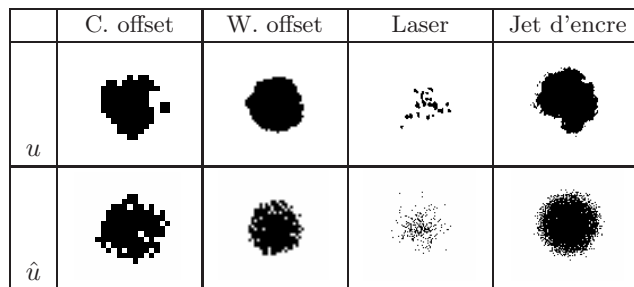
6 Conclusion

Nous avons développé un estimateur du modèle binaire spatial basé sur des noyaux Gaussien généralisés et avons proposé une méthode d’estimation de la borne de Cramer Rao a posteriori. Les performances de l’algorithme ont été étudiées sur des données simulées, les résultats montrent d’une part que l’estimateur offrent de bonnes performances et que la difficulté majeure est l’estimation du paramètre de forme β du noyau Gaussian généralisé. Sur des données obtenues par des imprimantes de différentes technologies, la qualité de la modélisation couplée à l’estimation montre la pertinence de notre approche, en particulier pour des impressions laser.

Références

[1] T. Kanungo, R.M. Haralick, H.S. Baird, W. Stuezle, and D. Madigan, “A statistical, nonparametric me-

TABLE 4 – Impression issues d’imprimante offset conventionnel, offset waterless, laser et jet d’encre u et simulation du modèle estimé \hat{u} .



thodology for document degradation model validation,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 22, no. 11, pp. 1209–1223, 2000.

- [2] H.S Baird, “The state of the art of document image degradation modelling,” in *Digital Document Processing*, pp. 261–279. Springer, 2007.
- [3] A.T. Phan Ho, B.A. Mai Hoang, W. Sawaya, and P. Bas, “Document authentication using graphical codes : impacts of the channel model,” in *Proceedings of the first ACM workshop on Information hiding and multimedia security*. ACM, 2013, pp. 87–94.
- [4] C. Baras and F. Cayre, “Towards a realistic channel model for security analysis of authentication using graphical codes,” in *Information Forensics and Security (WIFS), 2013 IEEE International Workshop on*. IEEE, 2013, pp. 115–119.
- [5] T.Q Nguyen, Y. Delignon, L. Chagas, and F. Septier, “Printer identification from micro-metric scale printing,” in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 6236–6239.
- [6] D. Arthur and booktitle=Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms pages=1027–1035 year=2007 organization=Society for Industrial and Applied Mathematics Vassilvitskii, S., “k-means++ : The advantages of careful seeding,” .
- [7] H. L. Van Trees, *Detection, estimation, and modulation theory*, John Wiley & Sons, 2004.