Indices de textures : application au classement de noyaux de cellules.

Guillaume Thibault — Bernard Fertil — Jean Sequeira — Jean-Luc Mari

Laboratoire LSIS, UMR CNRS 6168, équipe I&M
ESIL, case 925, 163 avenue de Luminy
13288 Marseille Cedex 9

Guillaume.Thibault@univmed.fr

RÉSUMÉ. *Dans cet article, il est présenté une étude sur la caractérisation et le classement de textures, effectuée à l'aide d'un ensemble de valeurs obtenues par le calcul d'indices. Pour calculer ces indices, il est nécessaire d'extraire de la texture un ensemble de données à l'aide de deux techniques : la construction de matrices constituant des représentations statistiques de la texture et le calcul de "mesures". Ces matrices et mesures sont utilisées comme paramètres de fonctions apportant des valeurs scalaires ou discrètes qui nous renseignent sur les caractéristiques de la texture. Toutes ces informations numériques permettent de construire un modèle qui décrit la texture et qui peut être ensuite utilisé par exemple pour la classifier. Une application est proposée pour diagnostiquer des noyaux de cellules chez des patients atteints de Progéria.*

ABSTRACT. *In this paper, we present a study on the characterization and the classification of textures. This study is performed using a set of values obtained by the computation of indexes. To get these indexes, we extract a set of data with two techniques: the computation of matrices which are statistical representations of the texture and the computation of "measures". These matrices and measures are subsequently used as parameters of a function bringing real or discrete values which give information about texture features. A model of texture characterization is built based on this numerical information, for example to classify textures. An application is proposed to classify cells nuclei in order to diagnose patients affected by the Progeria disease.*

MOTS-CLÉS : *Indices de formes et de textures, matrice de longueur de segments, matrice de surface de zones, classement.*

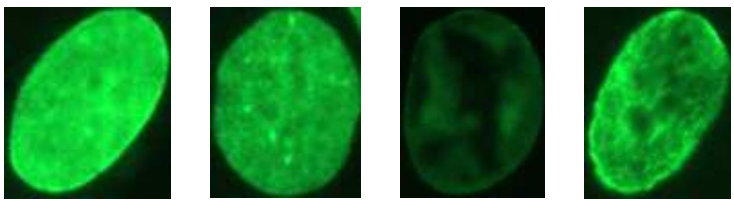
KEYWORDS: *Shape and texture indexes, gray level run length matrix, gray level size zone matrix, classification.*

1. Introduction

La reconnaissance de formes est une partie majeure de l'intelligence artificielle visant à automatiser le discernement de situations typiques au niveau de la perception. Elle est un enjeu majeur pour de très nombreuses applications : la reconnaissance des caractères manuscrits (numérisation des livres, lecture automatique des lettres postales et des chèques bancaires, etc.), la vidéo surveillance (reconnaissance faciale), l'imagerie médicale (échographie, scanner, imagerie par résonance magnétique), la télédétection, etc. Au cœur de la reconnaissance de formes, il y a une première étape incontournable : la caractérisation de formes. En effet, afin de pouvoir reconnaître un objet ou un individu, il faut tout d'abord le décrire et donc définir ses caractéristiques (morphologiques, géométriques, texturelles, etc.), puis retrouver et identifier ces mêmes caractéristiques sur la source numérique à analyser. Pour cela il est souvent nécessaire d'étudier les objets selon deux critères : le premier est la forme avec des méthodes d'analyse globale ou de contours et le deuxième la texture.

Le but de nos travaux est de créer un modèle afin de classer des noyaux de cellules de patients atteints par la maladie de la Progéria (également connue sous le nom de syndrome de Hutchinson-Gilford) dans les classes *noyau sain* ou *noyau pathologique*. Cette maladie orpheline (une centaine de cas dans le monde) de type laminopathie (Giovannoli *et al.*, 2003) provoque un vieillissement accéléré du patient. Nous disposons d'un ensemble de noyaux de cellules prélevés chez des patients, ainsi que d'une expertise de ces noyaux qui a révélé qu'il est nécessaire de traiter le problème suivant deux aspects : la forme du noyau (normale ou bien *boursouflée*) qui joue un rôle principal dans le diagnostic et l'homogénéité de la texture du noyau (texture *homogène* ou *non homogène*).

Notre contribution apporte une nouvelle méthode de caractérisation de l'homogénéité de la texture, basée sur la construction et l'analyse de matrices statistiques représentatives de la texture. Pour observer la texture, les images de noyaux sont acquises à l'aide d'un microscope à fluorescence et un marqueur de type FITC (Fluoresceine Iso Thio Cyanate) est employé afin de voir la répartition des lamines A et C (protéines codant pour la structure du noyau) qui sont utilisées par les experts pour le diagnostic des noyaux.



(a) (b) (c) (d)

Figure 1. *Exemples de noyaux marqués avec le FITC : à gauche deux noyaux à texture homogène (a, b) et à droite deux noyaux avec une texture non homogène (c, d).*

2. Travaux antérieurs

Les premiers travaux de caractérisation et de classement de la forme des noyaux (Wolberg *et al.*, 1995, Thibault *et al.*, 2008) avaient permis d'obtenir un taux de classement supérieur à 95% grâce à l'utilisation et à la création d'indices de formes dédiés. En revanche, les deux méthodes de caractérisation de la texture explorées ne furent pas totalement satisfaisantes. La première approche, basée sur les matrices de cooccurrences et les caractéristiques Haralick (Haralick *et al.*, 1973) avait obtenu un taux de classement de 90%. Une approche originale (Chen *et al.*, 1995) fondée sur des indices initialement conçus pour caractériser les formes ne permis pas d'atteindre un taux de classement supérieur à 85%. Ce taux est inférieur au taux de répétabilité des experts, c'est-à-dire au pourcentage de noyaux classés de la même façon par un expert lors de deux expertises successives. Bien que le résultat de la première approche soit supérieur au taux de répétabilité, nous avons souhaité parvenir à un résultat proche de celui obtenu pour la forme.

Pour cela, nous présentons les méthodes de classement et de validation employées tout au long de cet article, puis les techniques utilisées afin d'améliorer les résultats de classement de la texture : la notion de matrice de longueur de segment (*run length matrix*) est présentée, puis logiquement modifiée afin de créer une nouvelle méthode de caractérisation de l'homogénéité d'une texture. Ces techniques sont étudiées et validées dans le modèle mis en œuvre pour résoudre le sous problème de caractérisation de la texture.

3. Classement

Le but du classement est d'associer à chaque individu étudié une classe d'appartenance. Dans le sous problème qui nous concerne, l'objectif est de déterminer si un noyau de cellule a une texture "*normale*" (homogène) ou "*anormale*" (non homogène). Les méthodes de classement sont dites supervisées car elles font intervenir une expertise de référence. Dans notre cas, nous bénéficions de l'expertise des biologistes et généticiens qui ont déterminé les classes (*sain* et *pathologique*) et les sous classes (forme *normale* et *boursouflée*, texture *homogène* et *non homogène*, etc.)

Un modèle pour le classement est souvent construit par apprentissage, à l'aide de données dont on connaît la classe *a priori*. Bien que s'appliquant à un problème spécifique, le modèle doit être capable de généraliser (au sens des données). Pour cela, les données sont séparées en deux groupes : un échantillon d'apprentissage et un échantillon de validation. Le classificateur doit garder les mêmes performances lors de l'apprentissage et de la validation. Pour mettre en œuvre une méthode de classement, il faut au préalable construire un vecteur caractéristique de l'individu étudié. Le vecteur doit être pertinent au problème posé afin de permettre un bon classement et une bonne prédiction. Le risque majeur lorsque l'on fournit trop de caractéristiques au classificateur est l'apprentissage *par cœur*. Plus la dimension du vecteur est grande, plus le modèle sera adaptable et donc plus le classement sera bon, mais plus la validation du

modèle à l'aide d'individus non utilisés dans la phase de construction du modèle sera mauvaise. Il faut alors systématiquement valider chaque modèle construit et obtenir le meilleur taux de classement sur l'échantillon de validation.

Nous disposons de 2800 noyaux expertisés, dont seulement 135 possèdent un texture non homogène, soit environ 5%. Ce déséquilibre entre les deux classes *homogène* et *non homogène* ne permet pas de séparer les noyaux en deux échantillons équilibrés et représentatifs. Pour cela nous effectuons un *Down-Sizing*, c'est-à-dire que nous réduisons la taille de la classe la plus grande en sélectionnant les 135 noyaux les plus représentatifs (les parangons) parmi les noyaux à texture homogène qui sont ajoutés aux 135 noyaux à texture non homogène afin de construire l'échantillon de travail. Cette sélection s'effectue à l'aide d'une classification par k-moyennes (Hartigan *et al.,* 1979) qui est rendue robuste par le calcul préalable des formes fortes obtenues par la mise en œuvre successive de trois k-moyennes. Les parangons sont les noyaux les plus proches du barycentre de chaque classe. Nous aurions pu effectuer un *Over-Samping*, en sélectionnant aléatoirement des noyaux dans la classe non homogène et en les dupliquant afin d'équilibrer les classes, mais il a été montré dans (Japkowicz, 2000) que cette technique est généralement moins efficace. Par contre, nous prévoyons de mettre en oeuvre une méthode fondée sur la mesure d'entropie asymétrique (Marcellin *et al.,* 2006) qui présente l'avantage de prendre en compte la totalité des données disponibles. En raison de la faible taille de l'échantillon (270 individus), la validation sera systématiquement effectuée à l'aide d'un protocole *leave one out* (Martens *et al.,* 1998) : de manière itérative, on construit un modèle à partir de tous les individus disponibles sauf un, ce dernier étant utilisé pour évaluer le modèle. La répétition de cette opération pour tous les individus de l'échantillon fourni le taux de classement de la méthode.

La méthode de classement choisie pour le modèle est la *régression logistique* (Hosmer *et al.,* 1989). C'est un modèle linéaire particulièrement adapté pour les problèmes de classement à deux classes : $P = P(Y/\vec{x}) = \frac{e^{f(\vec{x})}}{1+e^{f(\vec{x})}}$ avec $\vec{x} = (x_1, ...x_n)$ le vecteur caractéristique de la donnée en entrée, $f(\vec{x}) = \sum_i \alpha_i x_i$ et $P(Y/\vec{x})$ la probabilité conditionnelle P de la variable $\vec{x}$ d'appartenir à la classe Y. Pour estimer les coefficients α_i du modèle, on utilise le plus souvent la méthode du maximum de vraisemblance qui maximise la probabilité d'obtenir les valeurs observées sur les échantillons de l'ensemble d'apprentissage. Elle consiste à rechercher les paramètres qui optimisent la fonction de vraisemblance $\pounds(\alpha, Y) = P^Y[1-P]^{1-Y}$. La régression logistique a été préférée à l'analyse discriminante (Fisher, 1936) pour sa plus grande fiabilité, souplesse, le peu de restriction sur les variables et l'explicité de ses résultats. D'une manière générale, le faible nombre de données disponibles, couplé à la qualité des résultats obtenus pour la forme et ceux présentés dans cet article pour la texture ne nous incitent pas à l'utilisation de méthodes non linéaires plus complexes comme les réseaux de neurones (McCulloch *et al.,* 1943), car les conditions de généralisation des résultats obtenus seraient difficiles à obtenir et vérifier.

4. Gray level run length matrix

La matrice de longueur de segments (*gray level run length matrix*) est une méthode statistique de caractérisation de la texture (Haralick *et al.,* 1973, Galloway, 1975, Chu *et al.,* 1990). Cette méthode effectue le comptage du nombre de segments de pixels de même intensité dans une direction donnée et les résultats sont représentés dans une matrice. Pour cela, une direction ($0°$, $45°$, $90°$ ou $135°$) et un nombre de niveaux de gris sont préalablement fixés. La valeur contenue dans la case (l, n) de la matrice est égale au nombre de segments de longueur l et de niveaux de gris n. Donc le nombre de colonnes de la matrice est dynamique car elle dépend de la longueur du plus long segment. De par sa conception, le calcul est symétrique, il est par conséquent inutile de la calculer dans les quatre directions complémentaires ($180°$, $225°$, $270°$ ou $315°$, on considère ici huit directions possibles entre le pixel étudié et ses voisins). La figure 2 montre un exemple du calcul de la matrice :

Image

1	2	3	4
1	3	4	4
3	2	2	2
4	1	4	1

$\Longrightarrow$

Gray level	Run length (j)			
i	1	2	3	4
1	4	0	0	0
2	1	0	1	0
3	3	0	0	0
4	3	1	0	0

Figure 2. *Exemple de remplissage de la matrice "run length" pour une image* 4×4*, dans la direction* $0°$ *et pour quatre niveaux de gris.*

Une fois la matrice remplie, onze indices sont calculés (Xu *et al.,* 2004) afin de construire le vecteur caractéristique de la texture. Pour construire notre modèle, nous calculons ces caractéristiques pour un niveaux de gris fixé et dans les quatre directions. Puis pour chaque indice, on calcule la moyenne de ses valeurs dans les quatre directions. Une recherche exhaustive a montré que le meilleur modèle est obtenu pour un ensemble de sept indices avec 32 niveaux de gris. Le taux de classement obtenu est de $84,81\%$, ce qui est un résultat moins performant que celui obtenu avec la matrice de cooccurrences et les caractéristiques Haralick (90%).

5. Gray level size zone matrix

Une texture homogène possède de grandes zones de même intensité et non pas des segments dans une direction donnée. Pour tenir compte de cette remarque, notre contribution comptabilise toutes les tailles des zones de pixels de même niveau d'intensité dans une matrice. Cette dernière est construite sur le principe de la *Run Length Matrix* : la valeur de la case (s, n) de la matrice contient le nombre de zones de taille s et de niveau de gris n. La figure 3 montre un exemple de calcul de cette matrice, baptisée *size zone matrix*.

Image

1	2	3	4
1	3	4	4
3	2	2	2
4	1	4	1

$\Longrightarrow$

Gray level i	Size zone (j) 1	2	3	4
1	2	1	0	0
2	1	0	1	0
3	0	0	1	0
4	0	2	1	0

Figure 3. *Exemple de remplissage de la matrice "size zone" pour une image* 4×4 *pour quatre niveaux de gris.*

La matrice produite possède un nombre de lignes fixe égal au nombre de niveaux de gris et une nombre de colonnes dynamique qui dépend de la taille de la plus grande zone. Plus la texture est homogène, plus la matrice sera *large* et *creuse*. Cette matrice possède l'avantage de ne pas nécessiter de calcul dans plusieurs directions, qui sont remplacés par un étiquetage des différentes zones. En revanche, il faut toujours spécifier un nombre de niveaux de gris, mais cela rend le remplissage davantage robuste au bruit. Nous calculons ensuite les mêmes onze indices que pour la matrice *run length* pour 32 niveaux de gris. Le taux de classement pour les onze indices est de 91.11% ce qui est meilleur que toutes les techniques testées jusqu'alors. Toutefois en étudiant les données et les indices, il apparaît qu'un cas particulier de texture n'est pas correctement caractérisé : les noyaux possédant de grandes zones homogènes, mais avec de fortes variations d'intensité entres les zones (figure 1 noyau *c*), ce qui en font des noyaux à texture non homogène. Pour caractériser ce type de noyaux, nous introduisons deux nouveaux indices qui sont des variances pondérées des niveaux de gris ou des tailles des zones :

$$Var_N = \sqrt{\frac{1}{N*S}\sum_{n=1}^{N}\sum_{s=1}^{S}\left(n*M(n,s)-\mu_N\right)^2}\ ,\ \mu_N = \frac{1}{N*S}\sum_{n=1}^{N}\sum_{s=1}^{S} n*M(n,s)$$

$$Var_S = \sqrt{\frac{1}{N*S}\sum_{n=1}^{N}\sum_{s=1}^{S}\left(s*M(n,s)-\mu_S\right)^2}\ ,\ \mu_S = \frac{1}{N*S}\sum_{n=1}^{N}\sum_{s=1}^{S} s*M(n,s)$$

avec N et S les dimensions de la matrice et $M(n,s)$ l'élément de coordonnées (n,s) de la matrice. Plus la texture contient de grandes zones avec des écarts d'intensité importants, plus la valeur de l'indice Var_N est grande. Une texture homogène engendre une valeur faible de cette indice. Il en est de même pour l'indice Var_S avec les écarts entre les tailles des zones.

Grâce à l'utilisation de ces deux indices dans le modèle composé des onze indices les plus pertinents, le taux de classement est désormais de 94.07%. La validité de notre modèle peut être observée sur la figure 4 qui montre la distribution des probabilités engendrées par le modèle. La forte répartition sur les extrémités de l'histogramme et la quasi absence de cas ambigus (probabilité autour de la valeur de décision 0.5), montrent l'efficacité de classement et la pertinence dans le choix des indices.

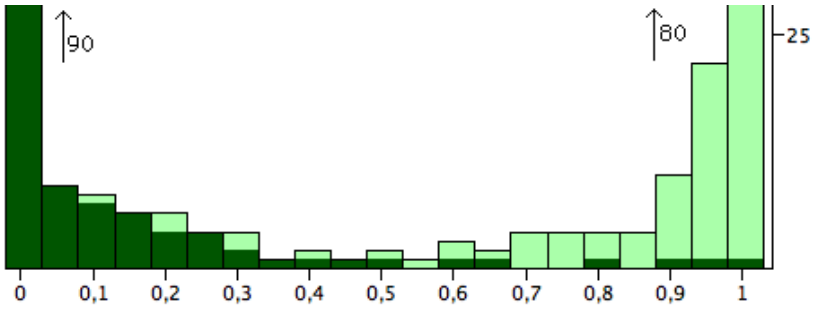


Figure 4. *Distribution des probabilités de classement données par le modèle : vert foncé (resp. vert clair) les noyaux à texture non homogène (resp. homogène). Plus la probabilité est proche de* 1 *(resp.* 0*), plus la texture est homogène (resp. non homogène). La majorité des noyaux sont classés avec des probabilités proches des extrêmes.*

6. Conclusions et travaux futurs

Dans cet article, nous avons présenté un problème de classement de la texture, appliqué au classement de noyaux de cellules. L'enjeu principal était de parvenir à caractériser de manière pertinente l'homogénéité de la texture. Pour cela, trois méthodes existantes de caractérisation de la texture ont été évoquées ou présentées : matrice de cooccurrences avec les caractéristiques Haralick, indices de formes pour la texture et matrice de longueur de segments. Ces méthodes n'ont pas permis de répondre de manière satisfaisante au problème (90%). Pour cette raison, nous avons proposé une nouvelle méthode de caractérisation de l'homogénéité d'une texture. Notre contribution procède à un étiquetage, puis à un dénombrement des tailles des zones de même niveau d'intensité. Ce dénombrement permet de construire une matrice représentative de l'homogénéité de la texture. Afin d'améliorer la pertinence de la méthode, nous avons proposé deux nouveaux indices de caractérisation de texture. Cette nouvelle approche permet d'obtenir un taux de classement de 94% de la texture des noyaux.

Le problème initial de notre travail était de classer les noyaux dans les classes *sains* et *pathologiques*. Notre contribution présentée dans cet article obtient un taux de classement de 86% de ce problème. De même, notre modèle de caractérisation de la forme présenté dans (Thibault *et al.*, 2008) apporte un taux de classement de 88.9%. En combinant ces deux modèles, nous obtenons 90.1% soit un gain de plus de 1% pour l'objectif initial. Bien que faible, cette amélioration est la meilleure que l'on puisse obtenir. En effet, on peut remarquer que parmi les 135 noyaux à texture non homogène, plus d'une centaine possèdent une forme anormale et sont donc classés par le modèle de forme. Notre contribution permet le classement de la trentaine de noyaux restant, soit environ 1% de la population. Les 6% des noyaux restants nécessitent d'utiliser d'autres caractéristiques et ne peuvent pas être classés par nos modèles.

Seulement 94% des noyaux peuvent être classés par leur forme ou/et leur texture. Dans nos travaux futurs, il nous faudra par conséquent proposer des modèles complémentaires afin de caractériser les critères occasionnels et ainsi améliorer davantage le taux de classement du modèle final.

7. Bibliographie

Chen Y. Q., Dixon M. S., Thomas D. W., « Statistical Geometrical Features For Texture Classification », *Pattern Recognition*, vol. 28, n° 4, p. 537-552, 1995.

Chu A., Sehgal C. M., Greenleaf J. F., « Use of gray value distribution of run lengths for texture analysis », *Pattern Recognition Letters*, vol. 11, n° 6, p. 415-419, 1990.

Fisher R. A., « The use of multiple measurements in taxonomic problems », *Annals of Eugenics*, vol. 7, p. 179-188, 1936.

Galloway M. M., « Texture analysis using grey level run lengths », *Computer Graphics Image Processing*, vol. 4, p. 172-179, July, 1975.

Giovannoli A. D. S., Bernard R., Cau P., Navarro C., Amiel J., Boccaccio I., Lyonnet S., Stewart C. L., Munnich A., Merrer M. L., Levy N., « Lamin A truncation in progeria », *Science*, vol. 300, n° 5628, p. 2055, 2003.

Haralick R. M., Shanmugam K., Dinstein I., « Textural features for image classification », *IEEE Transactions on Systems, Man and Cybernetics*, vol. 3, p. 610-621, 1973.

Hartigan J. A., Wong M. A., « A K-Means Clustering Algorithm », *Applied Statistics*, vol. 28, n° 1, p. 100-108, 1979.

Hosmer D., Lemeshow S., *Applied Logistic Regression*, John Wiley & Sons, Toronto, 1989.

Japkowicz N., Learning from Imbalanced Data Sets : A Comparison of Various Strategies, Technical report, AAAI Workshop on Learning from Imbalanced Data Sets, 2000.

Marcellin S., Zighed D.-A., Ritschard G., « An asymmetric entropy measure for decision trees », *Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU)*, p. 1292-1299, 2006.

Martens H., Dardenne P., « Validation and verification of regression in small data sets », *Chemometrics and intelligent laboratory systems*, vol. 44, n° 1-2, p. 99-121, 1998.

McCulloch W. S., Pitts W., « A logical calculus of the ideas immanent in nervous activity », *Bulletin of Mathematical Biophysics*, vol. 5, p. 115-133, 1943.

Thibault G., Devic C., Horn J.-F., Fertil B., Sequeira J., Mari J.-L., « Classification of cell nuclei using shape and texture indexes », *International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision (WSCG)*, Plzen, République Tchèque, p. 25-28, February, 2008.

Wolberg W., Street W., Heisey D., Mangasarian O., « Computer-Derived Nuclear Features Distinguish Malignant From Benign Breast Cytology », *Human Pathology*, vol. 26, Elsevier, New York, NY, United States, p. 792-796, July, 1995.

Xu D., Kurani A., Furst J., Raicu D., « Run-Length Encoding For Volumetric Texture », *International Conference on Visualization, Imaging and Image Processing (VIIP)*, p. 452-458, 2004.