Poster: Classification of cell nuclei using shape and texture indexes

Guillaume Thibault[1], Caroline Devic[2], Jean-François Horn[2],
Bernard Fertil[1], Jean Sequeira[1] and Jean-Luc Mari[1]
[1] LSIS Laboratory, Aix-Marseille II University, France
[2] Inserm U. 678, Paris VI University, France
Thibault@univmed.fr

ABSTRACT

In this paper, we present a study on the characterization and the classification of binary digital objects. This study is performed using a set of values obtained by the computation of "shape and texture indexes". To get the shape indexes, we extract a set of data called "measures" from 2D shapes, like for example surface and perimeter. These indexes are then used as parameters of a function returning a real value that gives information about geometrical and morphological features of the shape to analyze. A model characterizing the shape (and the texture) of objects is subsequently built. An application to the classification of cell nuclei (in order to diagnose patients affected by the Progeria syndrome) is proposed.

Keywords: Pattern recognition, shape and textures indexes, Haralick's features, cell nuclei classification.

1 INTRODUCTION

Pattern recognition is a major part of artificial intelligence that aims to automate the identification of typical situations. It is a major objective for many applications: handwritten character recognition (optical character recognition, etc.), video surveillance (facial recognition), etc.

At the heart of the pattern recognition issue, there is a first and unavoidable step: object characterization. It is often helpful to distinguish 2 classes of characteristics: the shape (polar signature [THK06], projection histograms [SR04], multi-scale curve smoothing for generalised pattern recognition (MSGPR) [KR06], etc.) and the texture [Har79]. Characterization of shape with shape indexes is more and more popular [IP97, TLG+03] espacially for learning-based classification [SEB+03]. Their flexibility, their simplicity of implementation and ease of use with a classifier make this approach an appropriate choice for many problems.

The aim of this paper is to create a model to classify blood cell nuclei in patients affected by Progeria syndrome. This rare syndrome is a laminopathy [GBC+03] that causes patients to age prematurely. To visualize nuclei, images are obtained using a fluorescent microscope that detects FITC tag (Fluoresceine IsoThioCyanate) showing the shape and the lamin A/C protein distribution.

WSCG 2008 conference proceedings, ISBN 80-903100-7-9
WSCG'2008, February 4 – February 7, 2008
Plzen, Czech Republic.

The first part of this paper defines the concepts of measures, shape indexes, examine their main properties and their usage for pattern recognition. Then the model used to solve the classification task is presented, studied and validated.

A set of over three thousand cell nuclei (figure 1) from patients with Progeria syndrome has been gathered. These nuclei were manually classified as *healthy* or *pathological*. The shape criterion was the most important diagnostic clue for 89% of the nuclei, but complementary information was obtained by a textural analysis relative to the homogeneity of the nucleus.

2 SHAPE INDEXES AND MEASURES

Shape indexes were presented for the first time in 1976 in the book by Santalo [San76] related to mathematical properties of convex shapes. The definition and properties of shape indexes can be found in [CC85, Fil95].

Shape indexes definition: *We call* shape indexes *parameters, coefficients or a combination of coefficients capable of providing numerical information about the shape of objects. A shape index must be dimensionless and invariant by translation, rotation and homothety.*

The majority of shape indexes is derived from an equality or an inequality observed on the shape being analysed. In [San76] the authors have established a set of inequalities about convex shapes in a continuous space: $P^2 - 4\pi A \geqslant \pi^2(\rho_e - \rho_i)^2$, with A the surface, P the perimeter, ρ_i (respectively ρ_e) the radius of the biggest (respectively smallest) inscribed (respectively circumscribed) sphere. All these inequalities use different shape parameters called "*measures*". The calculation of these measures is an unavoidable step for getting shape indexes (indexes are based on at least one

measure). Measures can occasionally possess dimension: two dimensions for the surface, no dimension for the number of holes. Measures without dimension are considered as shape indexes.
The ability to build indexes from inequalities inherent to the shape under investigation is a major advantage: every kind of shape can be characterized and classification tasks are facilitated.

2.1 Shape indexes and classification

The objective of this study is to differentiate *healthy* from *pathological* cell nuclei.
Classification methods are divided into two important families: supervised and unsupervised methods. Supervised methods are more powerful but require expert knowledge to learn from. In this study we benefit from the biologists' and geneticists' knowledge who have specified classes (healthy and pathological) and subclasses (*ellipsoidal* and *puffy* shapes, homogeneous and non-homogeneous textures), which has allowed us to use supervised methods.
The aim of classification methods is to build a classification model based on the data under investigation. Although being applied to a specific problem, the model must remain general within the framework of data. For this reason, the data are split into two sets: a learning set and a validation set. The classifier must have comparable performances for both the learning and the validation sets.
It is necessary to construct a characteristic vector for each data prior the classification phase. The vector must be relevant to the problem in order to allow accurate classification and prediction. The major risk when providing too many characteristics to the classifier is overfitting. The greater the vector's dimension is, the greater the flexibility of the model and the better the classification are, but the greater the likelihood that the model's performance will be poor for a data set not used during the validation is. It is therefore necessary to validate each model with respect to overfitting. In this study, the characteristic vector is composed of shape indexes and Haralick's features [Har79] for the texture.
Classification is achieved by the logistic regression [DS89]. It is a linear model particularly well adapted to classification problems with two classes. Logistic regression performs a statistical analysis on the learning set and uses a logical distribution function to predict a membership probability: $P = P(Y/x) = \frac{e^{f(x)}}{1+e^{f(x)}}$ with $x = (x_1, ...x_n)$ the characteristic vector of the initial data, $f(x) = \sum_i \alpha_i x_i$ and $P(Y/x)$ the conditional probability P of the variable x to belong to the class Y.

3 CHARACTERIZATION OF CELL NUCLEI WITH SHAPE INDEXES

Healthy nuclei have ellipsoidal shape while pathological nuclei are puffy so that concave border areas are visible. For this reason, a set of fourteen shape indexes was obtained from the scientific literature. It also appeared interesting to create three additional indexes specifically designed for this study.

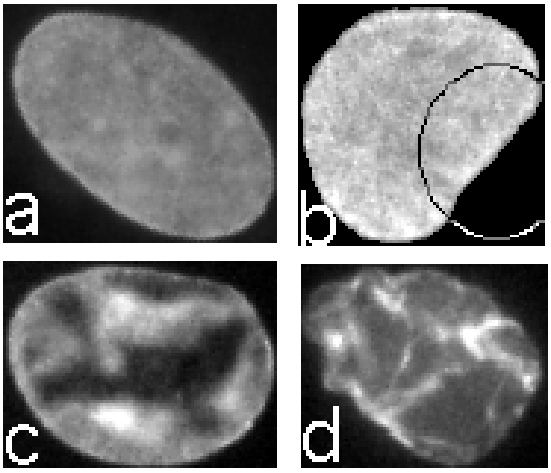

Figure 1: Four cell nuclei: *a* healthy, *b* puffy, *c* non homogeneous texture, *d* puffy and non homogeneous texture.

3.1 Three new shape indexes

Cell nuclei have a near elliptic shape when they are healthy. It is consequently judicious to build indexes characterizing the elliptic nature of the cells. The area of an ellipse A equals πab, with a the semi major axis and b the semi minor axis. Ellipses have some interesting properties that can be "measured": $R_{max} = a$ and $R_{min} = b$ or $R_{max} = \frac{1}{2}L_{AP1}$ and $R_{min} = \frac{1}{2}L_{AP2}$, with R_{max} (respectively R_{min}) the greatest (respectively smallest) radius and L_{AP1} (respectively L_{AP2}) the length of the principal (respectively secondary) axis. Two shape indexes can be derived based on these equalities:

$$\psi_{1\,ellipse} = \frac{\pi R_{min} R_{max}}{A}\ ,\ \psi_{2\,ellipse} = \frac{\pi}{4}\frac{L_{AP1}L_{AP2}}{A}$$

Denominators and numerators are equal in the case of an ellipsis and the index values are 1.
Pathological nuclei are currently not convex and consequently have concave border areas. To quantify these concave areas, it is possible to calculate the number of connected components N_{Cce} remaining when the shape is substracted from its convex hull. In the following, those connected components will be called "gap components". In this work, a normalized version of N_{Cce}, $\psi_{N_{Cce}}$, is used by the classifier:

$$N_{Cce} = card(ConvexHull(F)\backslash F)\ ,\ \psi_{N_{Cce}} = \frac{1}{1+N_{Cce}}$$

$\psi_{N_{Cce}}$ is equal to 1 if no gap component is found and tends towards 0 as the number of gap components in the shape increases.
In practice, the components the surface of which equals one pixel are due to resolution errors cannot be considered as gap components. In fact, even small gap components (i.e a few pixels) may not be significant at least with respect to the classification of nuclei. The size and number of gap components N_{Cce} must be taken into account when diagnosing nuclei elements. A systematic analysis of the percentage of correct classifications versus the minimum size and the number was conducted. The highest classification rate (90%) is obtained by considering that nuclei with one 32-pixel at least gap component or two 12-pixel at least gap components have abnormal shapes.

3.2 Construction and validation of the model for nuclei characterization

The model of classification that is used in this study build a linear combination of the indexes in order to predict the class. The efficiency of classification (on the validation test) relies on the selection of the best subset(s) of indexes.
Best all subsets research on the seventeen indexes is performed to find the best combination of indexes. For the validation, the *K-Fold* protocol is used (with $K = 10$). A subset composed of eleven indexes (see appendix A) yields the best classification rate. The clearly bimodal distribution of the classification probabilities of the nuclei associated with the high classification efficiency demonstrate the relevancy of the selected indexes (figure 2).

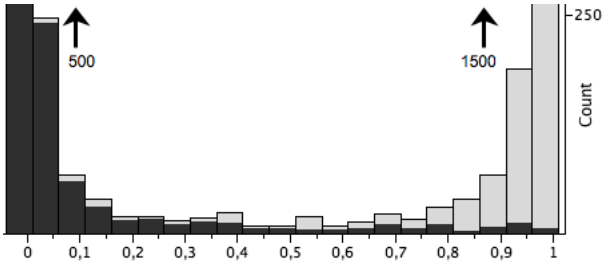

Figure 2: Distribution of the classification probabilities as given by the model to the nuclei from the validation set. The closer to one the probability, the more convex the nucleus. Dark areas and light grey areas are for puffy shapes and ellipsoidal shapes respectively.

4 TEXTURE CHARACTERIZATION

Although very good results have been obtained by modelizing shape, this only resulted in 89% of the global classification of nuclei into "healthy" or "pathological" groups. In order to improve this performance, it is necessary specifically analyse the homogeneity of the texture of the nuclei.

The lamin A/C distribution is homogeneous for healthy nuclei. However, the experts only consider nuclei as having non homogeneous texture when it is not "highly homogeneous" (figure 1).
In order to characterize texture, a co-occurrence matrix (32 grey levels) is built, out of fifteen Haralick's features are calculated [Har79]. One of these features is the homogeneity. Homogeneity is higher when the same pair of pixels is frequently found, as it is the case when there is an uniform area or a spatial periodicity.
Contrasting with the analysis of the shape, the analysis of the texture provides totally unballanced classes ("*homogeneous*" and "*non homogeneous*"), with approximately twenty times more nuclei in the homogeneous class. To efficiently build the model, the number of elements in each class must be roughly comparable. For this reason, the learning phase was carried out with all the nuclei from the non-homogeneous class (116 items) and an equal number of nuclei chosen by selecting prototypes in the homogeneous class with the *K-means* procedure ($K = 116$). The validation is realized according to the "Leave One Out"-protocol. Best all subsets research on the fifteen indexes is performed to find the best combination of indexes. Best subset is made of height indexes (listed in appendix B) which performs 90% of good classification of texture. In addition, the distribution of the probabilities is less constrasted than the distribution of the probability for the shape (figure 3). Several causes may be invoked to explain these differences: the main reason is the far lower number of nuclei belonging to the "non homogeneous" class which reduces the learning possibilities. The second reason is the noise introduced by using the tags, which reduces the reliability of the prediction.

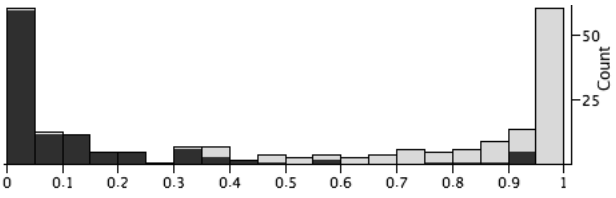

Figure 3: Distribution of the classification probabilities as given by the model to the nuclei from the validation set. The closer to one the probability, the more homogeneous the nucleus. Dark areas and light grey areas are for homogeneous and non homogeneous textures respectively

5 FULL MODEL: DIAGNOSIS OF NUCLEI

Two classification models have been built, characterizing the two main diagnostic parameters: the shape of the nucleus and its texture. These two models must be combined in order to establish the final model, capable of predicting the pathological aspect of nuclei. This

model takes advantage of the 11 shape indexes and the 8 texture model features. The classification success rate hence reaches 90% on the learning set and 89.5% on the validation set.

6 CONCLUSIONS AND PERSPECTIVES

In this study, we have presented a method for the classification of cell nuclei of patients affected by the Progeria syndrome. The first step was based on the study of the shape of cell nuclei using shape indexes. Appropriate indexes were specifically built. These indexes have subsequently shown the ability to correctly classify shape of nuclei with a success rate above 95%. With reliability and validation in mind, it is not planned to try and improve this result for two reasons: the first being that it would be necessary to introduce additional characteristics that would jeopardize learning (overfitting). The few tenths of a percent gained in correct classification would be lost in validation. The second reason is that the efficiency of this approach already matches the reproducibility rate of experts.

Next a model based on Haralick's features was built in order to handle the problem of texture characterization. This model has provided a satisfactory handling of the texture characterization (90% of good classification) and has allowed improving the final model. The shape model alone obtained a ratio of 88.9% correct classifications of the nuclei (healthy/pathological). The addition of the texture model allows an improvement in the diagnosis of less than 1%.

In light of these results, it seems necessary to improve the texture model. For this reason it is planned to extend the notion of shape indexes and to use it to analyze textures. A possible solution would be to consider a texture like an elevation map ; with each pixel no longer representing a greyscale but rather a altitude.

A SHAPES INDEXES

With F a form, A the surface, P the perimeter, B the barycenter, R_{max} (respectively R_{min}) the greatest (respectively smallest) radius, ρ_i (respectively ρ_e) the radius of the biggest (respectively smallest) inscribed (respectively circumscribed) sphere, L_{AP1} (respectively L_{AP2}) the length of the principal (respectively secondary) axis, D the diameter, E the thickness, N_{Cce} number of gap components.

$\text{Extension}_{\text{Diameter}} = \frac{E}{D}$, $\text{Extension}_{\text{Radius}} = \frac{\rho_i}{\rho_e}$

$\text{Circularity} = \frac{R_{min}}{R_{max}}$, $\text{Deficit} = 1 - \pi\frac{(\rho_e-\rho_i)^2}{P^2}$

$\text{Convexity}_{\text{Perimeter}} = \frac{P(ConvexHull(F))}{P(F)}$

$\text{Convexity}_{\text{Surface}} = \frac{A(F)}{A(ConvexHull(F))}$

$\text{Symmetry}_{\text{Besicovitch}} = \sup_{x\in F}\frac{A(F\cap Symmetric(F,x))}{A(F)}$

$\psi_{1\text{ Ellipsis}} = \frac{\pi R_{min}R_{max}}{A}$, $\psi_{2\text{ Ellipsis}} = \frac{\pi}{4}\frac{L_{AP1}L_{AP2}}{A}$

$\psi_{\text{N}_{\text{Cce}}} = \frac{1}{1+N_{Cce}} \in]0,1]$, $\psi_{2\text{ Parallelogram}} = \frac{A}{E\times D}$

B HARALICK FEATURES

With $p(x,y)$ the element (x,y) of the grey levels co-occurrence matrix, N the number of pixels of the texture to analyze, N_g the greyscale.

The standard deviation $\sigma = \sum_x\sum_y(p(x,y)-m)^2$

The correlation $\sum_x\sum_y(x-m)(y-m)p(x,y)/\sigma^2$

The average of the sums

The entropy of the sums

The entropy $\sum_x\sum_y p(x,y)log(p(x,y))$

The standard deviation of the differences

The homogeneity $\sum_x\sum_y\frac{1}{1+|x-y|}p(x,y)$

The dissimilarity $\sum_x\sum_y|x-y|p(x,y)$

REFERENCES

[CC85] Michel Coster and Jean-Louis Chermant. *Précis d'analyse d'images*. Editions du CNRS, 1985.

[DS89] Hosmer D.W. and Lemeshow S. *Applied Logistic Regression*. John Wiley & Sons, Toronto, 1989.

[Fil95] Isabelle Fillere. *Outils mathématiques pour la reconnaissance de formes*. PhD thesis, Université de St Etienne, Septembre 1995.

[GBC+03] Annachiara De Sandre Giovannoli, Rafaelle Bernard, Pierre Cau, Claire Navarro, Jeanne Amiel, Irene Boccaccio, Stanislas Lyonnet, Colin L. Stewart, Arnold Munnich, Martine Le Merrer, and Nicolas Levy. Lamin a truncation in progeria. *Science*, 300(5628):2055, 2003.

[Har79] R. M. Haralick. Statistical and structural approaches to texture. In *Proceedings of the IEEE*, volume 67, pages 786–804, 1979.

[IP97] Jukka Iivarinen and Markus Peura. Efficiency of simple shape descriptors. In *International Workshop on Visual Form*, pages 28–30, May 1997.

[KR06] Kidiyo Kpalma and Joseph Ronsin. Multiscale contour description for pattern recognition. In *IEEE Transactions on Pattern Recognition Letters*, volume 27, pages 1545–1559. Elsevier, October 2006.

[San76] L.A Santalo. *Integral Geometry and Geometric Probability*. Addison Wesley, 1976.

[SEB+03] A. Sboner, Claudio Eccher, E. Blanzieri, P. Bauer, Mario Cristofolini, G. Zumiani, and Stefano Forti. A multiple classifier system for early melanoma diagnosis. *Artificial Intelligence in Medicine*, 27(1):29–44, 2003.

[SR04] Hasan Soltanzadeh and Mohammad Rahmati. Recognition of persian handwritten digits using image profiles of multiple orientations. In *IEEE Transactions on Pattern Recognition Letters*, volume 25, pages 1569–1576. Elsevier, 2004.

[THK06] Abdelmalek Toumi, Brigitte Hoeltzener, and Ali Khenchaf. Classification des images ISAR pour la reconnaissance des cibles. In *XIIIème Rencontres de la Société Francophone de Classification (SFC)*, 2006.

[TLG+03] V. M. Tuset, I. J. Lozano, J. A. González, J. F. Pertusa, and M. M. García-Día. Shape indixes to identify regional differences in otolith morphology of comber. In *Journal of Applied Ichthyology*, volume 19, pages 88–93, April 2003.