Modélisation et classification pour la rétroconversion des documents

Pierre Héroux, Éric Trupin, Yves Lecourtier

MODÉLISATION ET CLASSIFICATION POUR LA RÉTROCONVERSION DES DOCUMENTS

Pierre Héroux, Éric Trupin et Yves Lecourtier

Laboratoire Perception, Systèmes et Information
UFR des Sciences et Techniques
Université de Rouen
76821 Mont−Saint−Aignan Cedex
Pierre.Heroux@univ−rouen.fr

RÉSUMÉ

Dans cet article, nous présentons une modélisation des documents orientée vers le processus de rétroconversion des documents papier vers un format électronique. Nous présentons également les méthodes de classification exploitant la modélisation des documents proposée. Ces méthodes de classification sont, pour les unes, utilisées pour effectuer la classification des éléments composant le document, pour les autres, utilisées dans la classification des documents eux−mêmes. Ces méthodes de classification sont combinées de façon à s'intégrer dans une stratégie globale de rétroconversion. Cette stratégie est basée sur un cycle d'interprétation au cours duquel différents niveaux d'interprétation interagissent. Ce cycle, appelé « cycle d'interprétation », est articulé autour de l'analyse et de la résolution d'incohérences dans la description des documents. Les premiers résultats concernant aussi bien la classification des objets constitutifs des documents que celle des documents eux−mêmes mettent en valeur l'utilisation du cycle d'interprétation.

Mots Clés : rétroconversion des documents, modélisation des documents, classification structurelle, cycle d'interprétation.

1. INTRODUCTION

Cet article traite de la rétroconversion des documents papier vers un format électronique. La rétroconversion consiste à construire une représentation des documents selon un formalisme précis, à partir de l'image du document.

L'intérêt porté à la rétroconversion des documents se justifie par le fait qu'elle permet de faire bénéficier les documents papier des avantages offerts par les représentations électroniques de haut niveau. En particulier, les documents ainsi convertis pourront être diffusés à une large échelle via les réseaux numériques. Ils pourront également être édités ou mis à jour plus aisément à l'aide de logiciels adaptés. Enfin, ils pourront être archivés sur support magnétique et indexés aussi bien sur leur aspect visuel que sur leur contenu textuel.

Le formalisme utilisé dans le processus de rétroconversion doit être aussi complet que possible de façon à ce que la représentation du document ainsi interprété puisse être facilement traduite vers un format électronique adéquat.

D'autre part, ce formalisme utilisé pour construire la représentation des documents doit être adapté au processus de rétroconversion. En effet, il doit

permettre la description de l'ensemble des niveaux d'interprétation du document, depuis le bas niveau vers les niveaux d'interprétation les plus hauts.

Il existe différentes vues du document [1]. En effet, un document peut être décrit par son image, par sa structure physique, sa structure logique [2]... La structure physique représente l'aspect visuel du document. Elle est composée d'objets physiques (objets composites ou objets de base) organisés de façon hiérarchique. Chaque objet physique représente une entité graphique ou matérielle du document (volume, page, colonne, bloc...). La figure 1 représente l'image d'un document ainsi que la structure physique associée. Les nœuds représentent les objets physiques, les arcs pleins représentent des relations de voisinage et les arcs pointillés des relations hiérarchiques composé/composants.

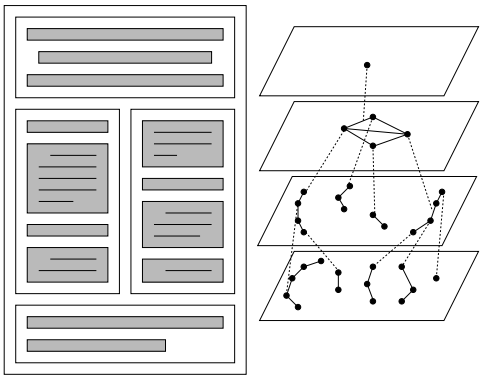

Figure 1 Modélisation physique d'une page d'un document synthétique

La structure logique représente l'organisation du document sur la base d'un découpage axé sur l'articulation de l'information véhiculée par le document. Cette structure, composée d'objets logiques, est elle aussi organisée de façon hiérarchique. La structure logique représente la façon dont le document peut être découpé en titre, sections, sous−sections, paragraphes... La figure 2 représente ce que pourrait être la structure logique du document de la figure 1.

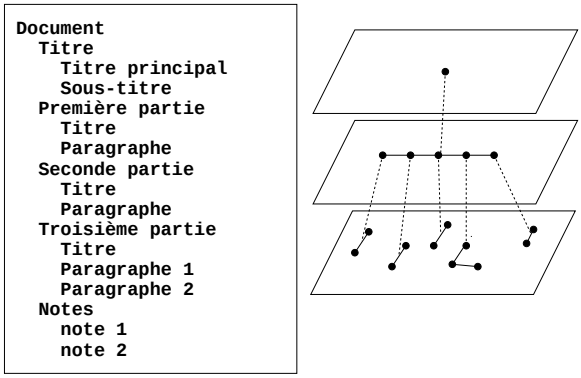

Figure 2 Modélisation logique d'un document synthétique

La phase d'extraction de la structure physique d'un document à partir de son image est appelée analyse de document. L'interprétation de document consiste à reconstruire la structure logique.

Une classe de documents est un ensemble de documents qui partage une partie commune de leur structure physique et de leur structure logique. La partie commune de la structure est appelée structure générique.

D'autre part, les objets peuvent eux aussi être regroupés en classes selon des caractéristiques communes. Une classe d'objets est décrite par un objet générique qui synthétise l'ensemble des caractéristiques communes à tous les objets de la classe.

Dans cet article nous présentons la modélisation et les méthodes de classification que nous utilisons au sein d'une stratégie d'interprétation et la rétroconversion des documents. Cette stratégie peut s'appliquer à la large gamme des documents qui peuvent être regroupés en classes et dont les objets constituants peuvent eux aussi être regroupés en classe. Cette stratégie est basée sur un cycle appelé « cycle d'interprétation ». Ce cycle inspiré par le « cycle perceptif » décrit par Ogier dans [3] fait interagir les phases d'analyse et d'interprétation du document. Chaque itération du cycle consiste à évaluer la pertinence de la description courante, à localiser les éventuelles incohérences et à proposer une nouvelle description dans laquelle les incohérences ont été traitées. Le cycle se termine lorsqu'il n'y a plus d'incohérence rencontrée lors de l'évaluation de la pertinence de la description.

Les éléments clés de cette stratégie sont d'une part, les méthodes de classification des objets constitutifs des documents et des structures physiques et logiques formées par l'organisation de ces objets, d'autre part la détection des incohérences ainsi que la façon dont ces incohérences sont résolues.

La section 2 décrit en détail la modélisation des documents employée ainsi que les méthodes utilisées pour effectuer la classification des objets et des structures. La section 3 présente des résultats ayant trait à la classification des documents et des objets qui les constituent. Enfin, la section 4 propose des perspectives découlant de l'examen des résultats des méthodes de classification en vue de l'implantation de la stratégie d'interprétation au sein d'un système de traitement automatique.

2. MODÉLISATION ET CLASSIFICATION

2.1 Représentation des objets

Dans cette section, nous décrivons les différentes méthodes utilisées dans la classification des objets, qu'ils soient physiques ou logiques, objets de base ou objets composites. Cette classification est effectuée en comparant l'objet spécifique à identifier avec l'ensemble des objets génériques représentant les classes d'objets.

Chaque objet contient trois informations de nature différente :

- un vecteur de caractéristiques numériques contient des informations propres à l'objet, indépendamment de son environnement ;
- un champ contient la classe de l'objet hiérarchiquement supérieur dans la structure ;
- un graphe représente la structure de l'objet.

Les nœuds du graphe représentent les classes des objets hiérarchiquement inférieurs. Les arcs du graphe représentent les relations de voisinages entre les objets fils.

Ces trois informations modélisent différents points de vue de l'objet considéré. Le vecteur de caractéristiques donne des informations intrinsèques à l'objet, indépendamment du contexte dans lequel il s'insère dans le document. Ce vecteur de caractéristiques est constitué des positions et dimensions de l'élément pour ce qui concerne les objets physiques, il est constitué d'informations ayant trait au formatage (alignement, taille, graisse...) pour les objets logiques. La classe de l'objet hiérarchiquement supérieur situe l'objet dans la structure. La structure de l'objet décrit non seulement la liste de ses constituants, mais également leur organisation, la façon dont ces constituants sont disposés les uns par rapport aux autres.

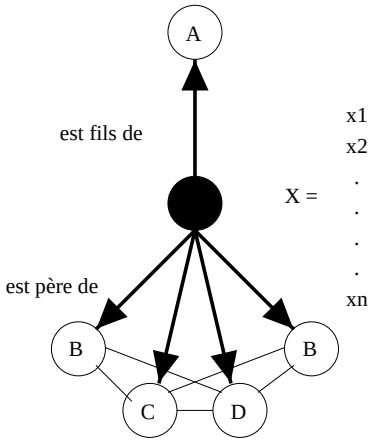

Figure 3 Structure d'un objet

2.2 Apprentissage − Synthèse des objets génériques

Les objets génériques décrivent une classe d'objets et synthétisent les informations communes à tous les objets de la classe. Ils sont générés grâce à un apprentissage supervisé et sont constitués d'un vecteur de caractéristiques, vecteur moyen de l'ensemble des vecteurs de caractéristiques des objets de la base d'apprentissage. La classe de l'objet hiérarchiquement supérieur attribuée à l'objet générique est celle la plus fréquemment rencontrée dans la base d'apprentissage. Le graphe représentant la structure de l'objet générique est le sous−graphe commun à tous les graphes représentant la structure des objets de la base d'apprentissage.

2.3 Classification des objets

La classification des objets consiste à identifier la classe à laquelle un objet inconnu appartient et donc à reconnaître l'élément décrit par cet objet. Cette classification est effectuée en trouvant lequel des objets génériques représentant les classes d'objets est le plus proche de l'objet à identifier. Les trois types d'information contenue dans la description de l'objet sont utilisées afin d'effectuer sa classification. Chacune d'elle donne une liste d'hypothèses valuées par une mesure caractérisant la confiance en l'hypothèse.

Une première méthode de classification exploite les propriétés propres à l'objet (vecteur de caractéristiques) en calculant la distance euclidienne entre le vecteur de caractéristiques de l'objet spécifique à identifier et les vecteurs de caractéristiques de chacun des objets génériques.

L'information concernant la classe de l'objet parent est également utilisée.

Enfin, une méthode de classification structurelle effectue la comparaison de la structure de l'objet avec celle de chacun des objets génériques. Chaque hypothèse est valuée par l'inverse d'une mesure de similarité qui est définie en déterminant le nombre de nœuds du sous−graphe commun aux deux graphes comparés [4]. Deux taux de recouvrement t_1 et t_2 sont définis comme le rapport du nombre de nœuds du sous−graphe sur le nombre de nœuds de chacun des graphes. La mesure de similarité *d* [5] peut alors être définie par

$$d = \frac{1}{t_1 \cdot t_2} - 1 \quad .$$

La classification définitive de l'objet est établie par la fusion des résultats fournis par les trois listes valuées d'hypothèses. Plusieurs méthodes peuvent sont envisagées afin d'effectuer une telle fusion.

Lors de la première itération du cycle de perception, lorsqu'aucun objet n'a encore été classifié, la seule méthode de classification prenant en compte les caractéristiques intrinsèques à l'objet, est utilisée puisqu'il n'est alors pas possible de faire référence à la classe des autres objets.

Finalement, lorsque tous les objets ont été classifiés, l'hypothèse retenue comme étant la classe à laquelle l'objet appartient est propagée vers l'objet dont il est un des composants et vers les objets qui le composent afin de mettre leur représentation à jour en vue de l'itération suivante.

2.4 Classification des structures de document

À chaque itération, une fois que tous les objets ont été classés, les structures physiques et logiques sont alors soumises à un classifieur structurel afin de déterminer la classe du document. La classification de chacune des structures (physique et logique) est établie en comparant la structure spécifique à identifier avec chacune des structures génériques des classes de documents connues. Finalement, une hypothèse concernant la classe du document est déterminée en retenant comme classe de document, l'hypothèse fournie par les classifications des structures si elles sont concordantes, ou celle donnant la meilleure mesure de confiance s'il y a conflit.

2.5 Cycle de classification

Un cycle de classification intègre la classification des objets et celles des structures. Ce cycle présenté sur la figure 4 a pour but de fournir une représentation stable du document.

À chaque itération du cycle, l'ensemble des objets physiques et logiques sont soumis à une classification.

Les structures sont à leur tour soumises à une classification. Le résultat de cette classification a pour but de proposer une hypothèse sur la classe du document.

Le résultat des classifications des objets et des structures est exploité lors de la phase de mise à jour de la représentation. En effet, ces résultats sont propagés dans la représentation afin de la rendre cohérente pour les classifications de l'itération suivante. Pour chaque objet, le résultat de la classification est propagé vers les objets qui lui sont reliés afin de mettre à jour à la fois leur structure et le champ relatif à la classe de l'objet parent. De la même façon, l'hypothèse sur la classe du document est exploitée afin de faciliter la classification.

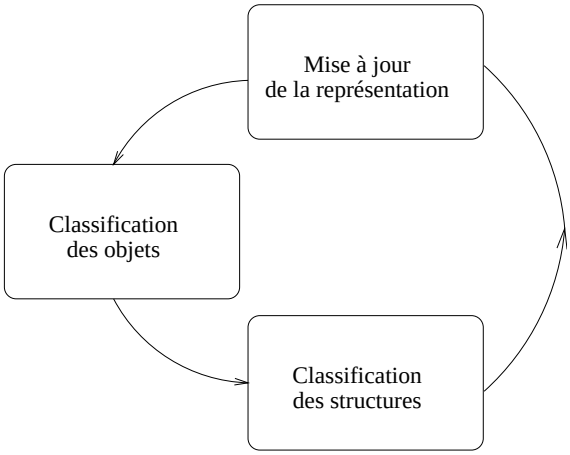

Figure 4 Cycle de classification

3. RÉSULTATS

Afin de valider notre approche, nous avons procédé à une expérimentation sur une base de documents synthétiques. Une interface graphique permet à un opérateur de saisir la représentation d'un document. À partir de la représentation d'un document, une base de documents est générée en appliquant diverses modifications sur la représentation. Ces modifications sont l'ajout d'un bruit gaussien sur le vecteur de caractéristiques de l'objet, la suppression d'objets dans la représentation et la modification d'étiquette associée aux objets. L'ampleur des modifications apportées par rapport au document original dépend de trois paramètres:

- l'ampleur du bruit gaussien ;
- le taux d'objets supprimés ;
- le taux d'objets affectés par une modification d'étiquette.

La figure 5 montre l'interface graphique utilisée pour la saisie, la visualisation et la modification de la représentation d'un document. La partie gauche représente la structure logique du document alors que la partie droite représente son aspect visuel.

Les figures 6 et 7 illustrent l'effet des modifications apportées par l'algorithme de génération de documents synthétiques, la figure 6 représentant l'aspect visuel du document saisi par l'opérateur et la figure 7 représentant celui d'un document généré automatiquement.

Cet algorithme de génération automatique d'une base de documents synthétiques est utilisé pour construire une base d'apprentissage de laquelle seront extraits tous les objets et les structures afin de construire les objets et les structures génériques représentant les classes d'objets et les classes de structures.

Cet algorithme est également utilisé pour construire une autre base de documents synthétiques qui servira de base de test.

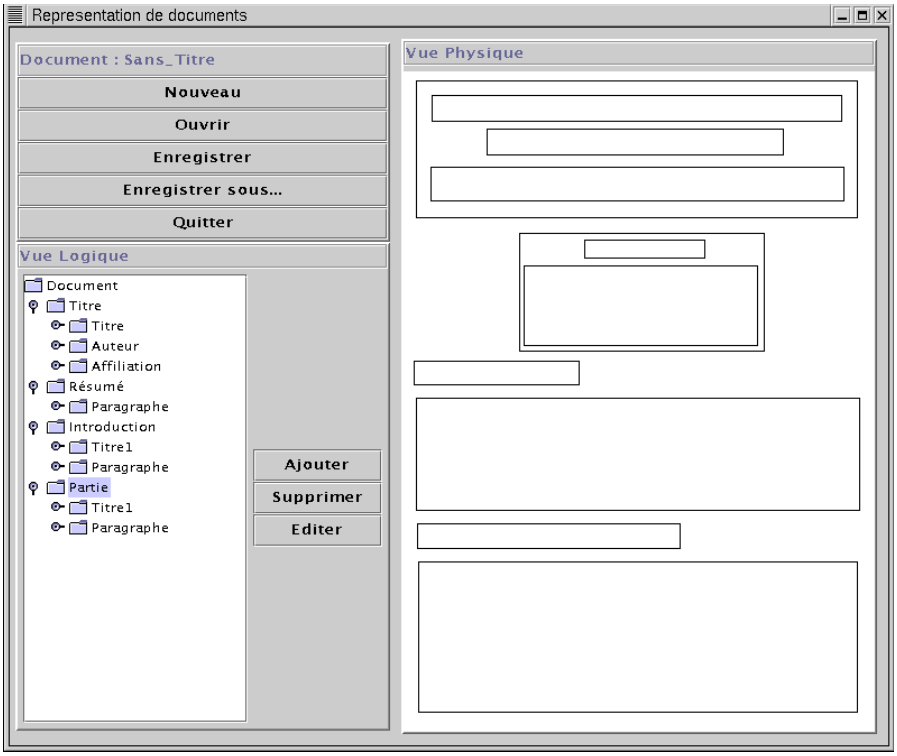

Figure 5 Interface graphique

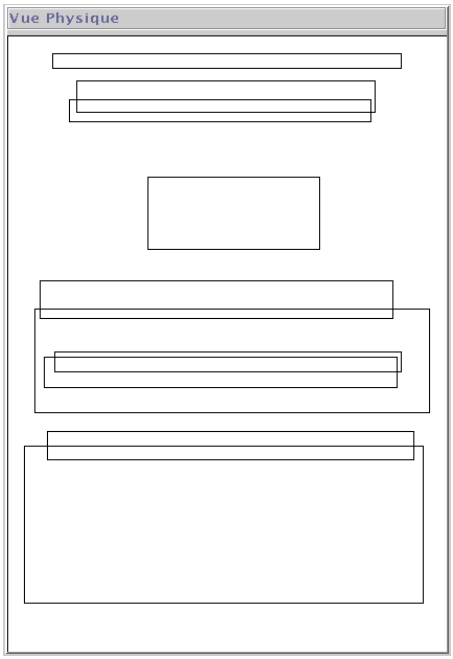

Figure 6 Aspect visuel d'un document synthétique saisi par l'opérateur

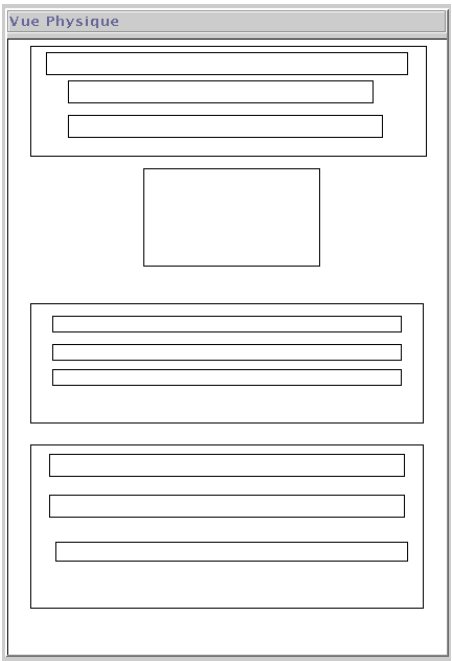

Figure 7 Aspect visuel d'un document synthétique généré automatiquement

Une première expérimentation a consisté à évaluer les différentes méthodes de classification des objets extraits des documents de la base de test. Cette classification a été effectuée sans utiliser le cycle d'interprétation. Les résultats obtenus montrent que seul le classifieur exploitant les caractéristiques intrinsèques de l'objet donne des résultats acceptables, le taux de reconnaissance variant entre 70 et 75% selon la taille de la base d'apprentissage que nous avons fait varier de 5 à 50 documents par classe avec 10 classes de documents, la base de test étant, pour sa part, constituée de 100 documents répartis également en 10 classes. L'extraction des objets de chacun des documents a constitué une base de test d'objets d'environ 1900 objets répartis en 18 classes.

Les résultats donnés par les classifieurs utilisant uniquement les informations structurelles ont donné des résultats bruts inférieurs à 30%. Cependant, il est à noter que le bruit ajouté afin de créer les documents synthétiques a induit des perturbations considérables sur les informations structurelles. En effet, par la variation de la position et de la dimension des éléments, certains objets physiques qui étaient liés du point de vue structurel ont vu cette liaison disparaître ou changer de nature. Il serait intéressant de considérer l'évolution du taux de reconnaissance des méthodes exploitant l'information structurelle lorsque le bruit ajouté aux vecteurs des caractéristiques intrinsèques a une influence moindre sur la structure du document.

Dans un second temps, plusieurs combinaisons de ces informations intrinsèques et structurelles ont été testées. Seule la méthode consistant à retenir, comme classe de l'objet à identifier, la classe présente en meilleur rang dans les listes fournies par les classifieurs a donné des résultats intéressants mais qui restent cependant du même ordre que la classification à partir du simple vecteur de caractéristiques.

En revanche, les informations de natures structurelles s'avèrent intéressantes lorsqu'elles sont utilisées en combinaison avec les informations intrinsèques au cours du cycle d'interprétation. En effet, nous avons vu les taux de reconnaissance s'améliorer au cours des itérations lorsque le résultats de classification des objets des itérations précédentes était propagé dans la représentation (cf. Tableau 8). Le fait que les taux de reconnaissance peuvent se dégrader sur deux itérations successives est du au fait que certaines interprétations étant stables, elles ne sont plus remises en cause. De ce fait, seul les documents les plus difficiles à reconnaître sont par la suite classifiés. Il est intéressant de remarquer que le taux de reconnaissance en fin de cycle, même s'il reste modeste, est supérieur à celui obtenu lors de la première itération.

Enfin, même si les résultats de la classification des objets représentant les éléments constitutifs des documents peuvent sembler modeste, il est à noter qu'ils semblent suffisant pour la reconnaissance de la structure des documents. En effet, nous avons constatés que dès lors que 50% des objets étaient reconnus les structures des documents elles reconnues à 100%. Bien que la base utilisée ne soit pas suffisamment conséquente, on peut tout de même expliquer ce phénomène par le fait que seul les objets les plus stables, et donc les plus faciles à reconnaître, sont pris en compte dans la structure générique.

Taille de la base d'apprentissage	Itération 1	Itération 2	Itération 3	Fin de cycle
5	52%	65%	62%	60%
10	49%	58%	55%	56%
20	46%	48%	56%	ne converge pas

Tableau 8 Évolution du taux de reconnaissance des objets

4. CONCLUSION

Dans cet article, nous avons présenté une modélisation adaptée au processus de rétroconversion ainsi que des méthodes de classification des objets représentants les éléments constitutifs des documents ainsi que des documents eux−mêmes. Les résultats obtenus à partir d'une base de documents synthétiques peuvent être selon divers point de vue. Tout d'abord, le taux de reconnaissance des objets peut paraître faible. Cependant, cette proportion d'objets reconnus suffit à identifier la classe de documents de façon fiable. D'autre part, s'il est vrai que les informations de type structurel n'apportent qu'une faible contribution à la classification, leur intérêt peut être appréciable. En effet, plutôt que de les faire intervenir systématiquement, ces informations pourraient être utilisées dans la résolution d'une incohérence localisée.

Enfin, on pourrait envisager d'améliorer le taux de reconnaissance en effectuant la recherche et la résolution d'un certain nombre d'incohérences dans la représentation.

5. BIBLIOGRAPHIE

[1] Stern, Y., Les quatre dimensions du document. *Document Numérique,* vol 1, 1997.

[2] International Standard Organization, *ISO 8613 : Information Processing. Text and Office System, Office Document Architecture (ODA) and Interchange Format,* 1989.[6] Diana, S., Trupin, E., Jouzel, F., Labiche, J., Lecourtier, Y., From Acquisition to Modelisation of a Form Base to Retrieve Information, *Fourth International Conference on Document Analysis and Recognition.* IAPR, 1997.

[3] Ogier, J−M., Mullot, R., Labiche, J., Lecourtier, Y., Interprétation de document par cycle « perceptifs » de construction d'objets cohérents. Application aux données cadastrales, *Traitement du Signal,* vol 12, 1995.

[4] Miclet, L., *Méthodes Structurelles pour la reconnaissance de formes.* Eyrolles, 1984.

[5] Héroux, P., Diana, S., Trupin, E., Lecourtier, Y., A Structural Classifier to Automatically Identify Form Classes. *Advances in Pattern Recognition, Lecture Notes in Computer Science,* vol 1451, 1998