QUATORZIEME COLLOQUE GRETSI - JUAN-LES-PINS - DU 13 AU 16 SEPTEMBRE 1993

INTERPRÉTATION AUTOMATIQUE DE DOCUMENTS APPLIQUÉE AU CAS DU CADASTRE FRANÇAIS

J.M. Ogier, R. Mullot, J. Labiche, Y. Lecourtier

La3i-LACIS,Université de Rouen
Place Emile Blondel, 76821 Mont Saint Aignan Cedex, FRANCE
Tél : 35.14.65.88 Fax : 35.14.63.49 e_mail : ogier@la3i.univ-rouen.fr

RÉSUMÉ

L'objectif de l'interprétation automatique de document est d'extraire le plus automatiquement possible l'information "pixellaire" afin de les stocker sous forme objet. Ce stockage nécessite une connaissance a priori sur la représentation graphique des éléments caractéristiques du document traité. Nous proposons dans cette communication un système d'interprétation de document appliqué au cas du cadastre français. Notre approche pour résoudre ce problème se déroule en deux phases. Dans un premier temps, nous appliquons un ensemble d'extracteurs de primitives graphiques bas niveaux. Dans un second temps, nous introduisons la connaissance a priori du document pour reconstruire les "objets" du cadastre en agençant les primitives extraites pendant la première phase.

ABSTRACT

The aim of the interpretation of documents is to extract as automatically as possible their characteristic elements in order to store them as object instead of pixel. This paper deals with a cadastral map interpretation device. Our approach to solve this problem is two fold : first, it consists in vectorizing the image by extracting the lowest information level. Then, using the low level primitives and introducing the knowledge of the cadastral map, it consists in reconstructing real cadatral entities. We present in this paper the different original tools allowing to extract the low level primitives.

I- INTRODUCTION

Le cadastre représente une vue macroscopique du territoire français. Chaque commune dispose de ses propres planches cadastrales comportant l'ensemble des informations sur la répartition des parcelles de terrain, leur numérotation, leur mitoyenneté ainsi que les parties communes telles que les routes, les voies privées.

Les nombreuses modifications, qui se comptent par millions pour une année, sont actuellement effectuées par les agents administratifs du cadastre français.

L'interprétation de document [1] [2] [3] [4] [5] vise à transformer l'ensemble des planches cadastrales acquises par un scanner, en informations de plus haut niveau : type de parcelles (habitation, cour, terrain), numéro de parcelle, segments constituant les contours. Outre le gain de place mémoire, ce type d'informations a un double avantage : tout d'abord, il permet d'extraire de chaque planche uniquement les informations utiles : les routes, les habitations, les terrains; Une sélection de l'information stockée est alors possible. De plus, il est très facile de modifier numériquement les informations détenues telles que les numéros ou les contours d'un objet. La fusion ou la réunion de deux parcelles ne posent alors aucun problème, et ne nécessite pas de redessiner l'ensemble de la planche manuellement.

A ce jour, de nombreuses municipalités, comprenant l'intérêt de tels outils, s'intéressent de très près aux différents produits permettant une interprétation automatique de documents.

Cette transformation de l'information comporte cependant de nombreuses difficultés. Le mode production en est une, puisque le cadastre français est réalisé manuellement. Il existe bien des règles d'écriture, mais elles sont plus ou moins bien suivies. Aussi, il n'est pas rare d'observer des hachures irrégulières, mais également des caractères dessinés différemment. De plus, il n'y a qu'un seul mode d'impression : le tracé noir. Tous les objets du cadastre sont représentés par la même couleur de tracé, voire la même épaisseur de trait. Il devient alors très difficile de différencier une hachure d'un "1", surtout si ces deux objets se touchent;

Notre démarche consiste à caractériser la spécificité de tel ou tel objet par certains attributs caractéristiques. Cette caractérisation peut, suivant le type d'objet manipulés, être locale, régionale, ou globale, ou une combinaison des trois approches.

Dans cet article, nous présenterons un certain nombre d'outils permettant l'extraction de certains objets et de ses attributs, en privilégiant la robustesse. D'autre part, nos outils répondront à une contrainte de précision. Le seuil de tolérance a été fixé au 1/10 de millimètre, ce qui peut paraître trop précis par rapport à la précision des planches réalisées manuellement. Dans le cas d'une acquisition au scanner à 300 dpi, cette tolérance équivaut à un résultat à + ou - 1 pixel.

Nous présenterons dans un premier temps, la méthodologie utilisée en vue de l'interprétation pour ensuite développer l'extraction des zones hachurées, puis des parcelles et des routes, enfin la vectorisation des contours des parcelles et le traitement des caractères en vue de leur reconnaissance.

II- METHODOLOGIE DE L'INTERPRETATION AUTOMATIQUE

Lorsque l'on analyse une planche cadastrale, on distingue essentiellement 3 objets différents :

- les parcelles hachurées : zones bâties,
- les parcelles non hachurées : zones non bâties,
- les routes.

Dans un premier temps, il semble difficile, en utilisant des critères simples, de discriminer les parcelles non bâties des routes. Ces deux objets comportent des caractères et sont non texturés. Par contre, les parcelles bâties peuvent être caractérisées par leur texture régulière, qui ne doit pas varier, a priori, d'une

parcelle à l'autre. Notre première étape consiste à extraire l'ensemble des objets composés de cette texture, sans perturber les autres attributs. Une fois les zones bâties étiquetées, on dispose donc d'un ensemble de parcelles et de routes. Les parcelles sont caractérisables par leur contour fermé, alors que les routes n'ont que des contours ouverts. Il est en effet difficile d'imaginer une route fermée à ses deux extrémités.

Dès lors, ils est possible d'utiliser un outil d'extraction des contours fermés. Cependant, les contours extraits et les segments de vectorisation ne sont pas identiques. Il est donc souhaitable de polygonaliser cette suite de points issus de l'extraction afin de ne recueillir que les segments les plus proches de ceux dessinés par l'agent administratif.

Enfin, il reste à traiter les caractères et les symboles contenus dans les parcelles et sur les routes afin de supprimer les éventuels bruits avant la phase de reconnaissance. Une fois ces traitements réalisés, on dispose donc des parcelles vectorisées, de ses attributs "caractères" et "type de zone". Il reste donc à mettre en place une stratégie permettant de reconstruire les parcelles composées de plusieurs contours fermés.

III- EXTRACTEUR DE ZONES HACHUREES

La hachure est un type de représentation très utilisée sur les documents techniques. Leur représentation est caractérisable par une suite de motifs réguliers. Il s'agit donc d'analyser une texture régulière. Une méthodologie employée consiste à caractériser cette texture, c'est à dire à trouver un ensemble de caractéristiques permettant la discrimination de la texture en présence, puis d'extraire de l'image, l'ensemble des zones comportant les caractéristiques extraites.

Les démarches de caractérisation et d'extraction varient suivant les auteurs, et suivant le type de texture.

Rappelons que dans notre cas, la texture est quasi régulière, c'est à dire que des variations peuvent être sensibles entre les hachures. D'autre part, il ne faut qu'en aucun cas que l'extracteur des hachures supprime des attributs. Il doit donc être très sûr.

III.1- Caractérisation de la texture régulière.

La caractérisation de la hachure suit les travaux de volet **[6]** qui modélise une texture régulière par une primitive de taille réduite. Dans notre étude, les primitives utilisées sont une succession d'occurrences blanches et noires dans le sens vertical et horizontal. A partir d'une zone prototype ne comportant que des hachures de bonne qualité, on calcule 4 histogrammes des fréquences :
- Occurrences blanches horizontales
- occurrences noires horizontales
- occurrences blanches verticales
- occurrences noires verticales.

Les histogrammes issus de cette analyse permettent de caractériser des occurrences optimales noires et blanches verticales et horizontales. La première primitive issue de cette technique est un masque composé de l'occurrence noire + occurrence blanche :

Cette première primitive permet de détecter localement les composantes de la texture. cependant, cette caractérisation doit être validée par une analyse plus globale. Le critère utilisé pour cette validation est l'enchaînement des primitives locales de la texture. Pour se faire, on utilise les histogramme des occurrences blanches et noires verticales afin de déterminer l'inclinaison globale des hachures. Un nombre N est alors déterminé définissant un nombre de ligne de validation de l'inclinaison globale de la hachure. Cette inclinaison est déterminée à travers le décalage existant entre les différents centres des primitives extraites sur des lignes successives. Le décalage est calculé à partir de la différence d'abscisse des centres des primitives de 2 lignes successives.

Le décalage global représente la différence cumulée d'abscisses entre N primitives successives verticales. Ce décalage, beaucoup plus fin dans son estimation , permet de caractériser l'inclinaison de la hachure.

A l'issue de cette phase de caractérisation de la hachure, on dispose de 4 paramètres :
- occurrence noire horizontale
- occurrence blanche horizontale -> constituant la primitive de base de la texture
- occurrence blanche verticale
- le décalage globale des centres des primitives sur une occurrence blanche verticale.

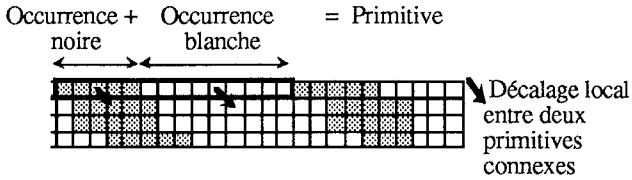


Figure 1 : Détermination du pattern ligne et de l'inclinaison locale des hachures.

III.2- extraction des zones hachurées

L'extraction des zones hachurées des plans cadastraux, basée sur la technique de caractérisation définie précédement, repose sur le concept de **"hachure potentielle"**. Toute "primitive" détectée dans l'image est étiquetée comme "hachure potentielle locale". Lors de la progression dans l'image, la validation de la présence de hachure réelle devient effective lorsque la propriété "hachure potentielle locale" a été propagée sur une plage suffisamment importante de lignes successives. La propagation de la "hachure potentielle locale" se fait si, sur deux lignes consécutives où la primitive peut être superposée, le décalage entre ces primitives est compatible avec le décalage local obtenu par **apprentissage**.

Ainsi, la validation de la propriété "hachure potentielle" en "hachure certaine" est obtenue si les conditions suivantes sont remplies :

- la propriété "hachure potentielle locale" s'est propagée sur un nombre suffisant de lignes. (Ce nombre N est obtenu par apprentissage et est égal à l'occurence verticale blanche).
- le décalage global entre la première et la dernière ligne de "hachure potentielle locale" est égal au décalage global obtenu par apprentissage.

Un étiquetage adaptatif permet d'appliquer cette propagation de la propriété "hachure potentielle". Suite à ce traitement, toutes les zones hachurées sont caractérisées par la même étiquette.

Une seconde partie consiste à extraire de l'image toutes les formes connexes étiquetées comme "hachure certaine". Les accidents contenus dans chaque zone hachurée (entre autres, le numéro identificateur des parcelles) sont également recensés et une liste des objets "batiment" est construite.

Les résultats illustrant l'application de cette méthode sur nos images sont présentés sur la figure 1.

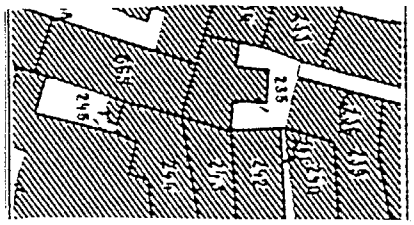


image source

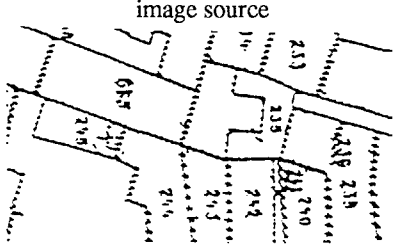


image résultat

Figure 2: résultat de l'extraction des zones hachurées

IV EXTRACTION DES ZONES DE CARACTERES.

IV.1 Localisation et orientation.

Après traitement de l'attribut hachure, les objets restant à extraire concernent toutes les zones textuelles et les objets linéaires. Nous présentons dans cette partie l'extraction et la

préclassification des composantes connexes, opération qui nous permet de localiser et de calculer l'orientation des caractères. En effet, afin de distinguer les attributs comme les caractères d'autres attributs (comme les symboles particuliers du cadastre), une préclassification des composantes connexes est réalisée. Cette préclassification est basée sur des critères géométriques (taille, rapport (périmètre/surface), ...) et sur des critères de localisation (il est quasi-impossible de rencontrer des symboles dans une parcelle par exemple). Sur une planche cadastrale, les caractères sont, soit des numéros identificateurs de parcelle, soit des noms identifiant des objets comme les routes ou les rivières. Les caractères sont généralement alignés mais peuvent présenter des orientations très variables. La localisation des caractères est obtenue à partir de deux sources d'information.

Tout d'abord, les "accidents" des différentes zones hachurées recensées sont analysés pour vérifier si ils correspondent effectivement à des zones de caractères. D'autre part, un extracteur de composantes connexes est déclenché sur toute l'image pour recenser les zones de caractères. C'est ensuite à ce niveau qu'une préclassification sur des critères géométriques s'effectue pour séparer la couche caractère de la couche symbole.

Après avoir localisé les caractères, leur orientation est calculée grâce à un couplage de deux algorithmes : le premier traitement consiste à regrouper dans une même zone de caractères les composantes géométriquement proches et de taille similaire ou les composantes localisés dans le même type d'objet (dans une parcelle par exemple). L'alignement de ces composantes est ensuite exploité pour évaluer de manière globale l' orientation de chacune des composantes (l'orientation des composantes est perpendiculaire à l'orientation de l'alignement du regroupement des composantes). Le calcul précis de l'orientation de ces composantes est ensuite affiné par un calcul d'axe principal d'inertie sur chacune des composantes. Les résultats illustrant l'application de ces algorithmes sont présentés sur la figure 3 :

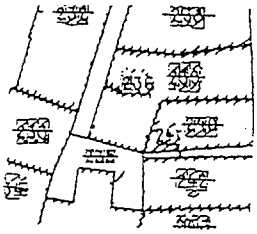


Figure 3 : Localisation des caractères et calcul de leur orientation

IV.2. Nettoyage et reconnaissance.

Après avoir localisé les différentes zones de caractères, nous disposons d'une base de données dans laquelle sont stockés les différents objets identifiés comme caractères. Cependant, certains d'entre eux nécessitent un "nettoyage" à cause de certaines "barbules" liées à des morceaux de hachures collées aux caractères. Avant de présenter nos caractères à un module de reconnaissance, nous avons développé un module de nettoyage original des caractères. Ce module de nettoyage utilise la connaissance a priori que l'on a sur ces barbules puisqu'il s'agit de hachures. Le jeu de paramètre utilisé pour la caractérisation des hachures est donc ici réutilisé pour nettoyer nos caractères. En fait, l'allure des barbules est modélisée suivant une loi binomiale qui nous permet de localiser parfaitement ces imperfections collées aux caractères. Les résultats concernant ce nettoyage des caractères sont présentés sur la figure 4.
Nos études portent actuellement sur l'extraction de primitives invariantes à la rotation pour reconnaitre les caractères. **[4]**

Image initiale Image résultat

Figure 4 : Nettoyage des caractères.

V. EXTRACTION DES OBJETS LINEAIRES.

V.1. Stratégie.

Après le traitement des zones hachurées et des caractères, l'information de base est séparée en différentes couches. En effet, les pixels sont de quatre types différents : Ils sont étiquetés comme "zone hachurée", comme caractères, comme fond ou comme objet autre (objet linéaire...). Dans ce paragraphe, nous présentons l'analyse des contours de parcelles (qui constituent la partie la plus importante des objets linéaires). Pour extraire de l'image tous les contours (qui peuvent correspondre à une parcelle ou à une construction), une extraction des composantes connexes étiquetées comme fond et zone hachurée est réalisée. En général, les composantes connexes extraites correspondent aux parcelles ou aux constructions. Des hypothèses sont alors émises pour établir la relation entre les contours des composantes connexes détectées et les limites des parcelles. Du point de vue de la reconnaissance de formes, cette approche est très intéressante car elle permet l'interprétation directe et la reconstruction des objets du cadastre.

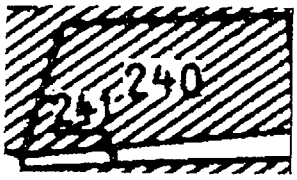
V.2. Vectorisation des contours des parcelles

Les contours des parcelles doivent être calculés avec précision. Dans ce paragraphe, nous présentons deux algorithmes différents qui permettent de réaliser la vectorisation des contours de parcelle. Nous présentons également une comparaison qualitative de ces méthodes.

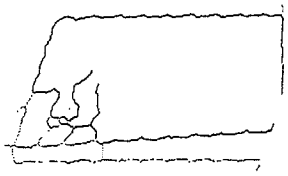
La première approche que nous avons utilisée pour vectoriser est la plus classique. La première étape de ce traitement classique des contours est une squelettisation **[10]** de l'image qui permet d'obtenir une image avec des lignes de un pixel de large. Puis, pour structurer l'information, la seconde étape exécute un suivi du squelette, traitement qui n'était pas réalisé durant la phase de squelettisation.

Au lieu de réaliser une extraction du squelette couteuse en temps, la deuxième méthode réalise l'extraction du contour de parcelle par un algorithme original de suivi de ligne. Cet algorithme est dérivé des travaux de [7] dans lesquels ce principe est utilisé pour l'extraction du réseau routier de cartes digitalisées.

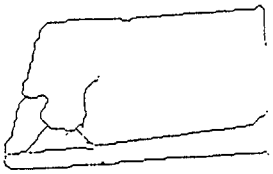
L'élément principal de cet algorithme est présenté ci-dessous : il consiste à suivre la ligne en progressant le long de son "axe" central. Cet axe est défini comme un ensemble de segments, chacun de ces segments étant obtenu à partir des 8 directions de Freeman. En fait, à chaque point P_i, la direction sélectionnée est la direction de Freeman qui permet le plus grand déplacement dans le trait sans rencontrer de transition de couleur (une transition de couleur correspondant à un bord du trait). De cette manière, chaque progression dans la ligne est une étape dont la longueur dépend de la largeur et de la courbure de la ligne considérée A chaque itération, si aucune anomalie n'a été détectée (épaisseur anormale de la ligne), le point atteint est recentré dans la ligne. Dans le cas d'une "surépaisseur", ce point singulier appelé nœud est mémorisé.
Après avoir suivi une ligne, les nœuds mémorisés sont analysés pour voir si ils correspondent à un bord discontinu (qui peut être soit une ligne réellement discontinue, soit un bruit de l'image) soit à un coin de la parcelle.
Les résultats concernant l'application de ces deux méthodes, présentés figure 5 sont intéressants car ils soulignent les avantages et les inconvénients de chaque méthode.



parcelle originale



parcelle squeletisée



parcelle obtenue avec l'algorithme de suivi de ligne

Figure 5: Comparaison des deux méthodes d'extraction de contours

L'application de ces méthodes nous amène à conclure que :

- la phase de squelettisation est une phase consommatrice de temps calcul, car elle nécessite plusieurs lectures (deux dans notre cas) de l'image et un suivi de squelette pour structurer l'information. La précision de la méthode est intéressante car elle respecte l'axe central réel de la ligne dans le cas où le segment est "propre".

- l'algorithme de suivi de ligne est plus efficace du point de vue du temps calcul car il analyse directement l'information significative (i e. l'objet à vectoriser : la ligne) sans traiter le reste de l'image. Par ailleurs, l'inconvénient principal de cette méthode est le manque de précision. En effet, comme la progression se fait par étape à l'intérieur de la ligne, le pseudo-squelette obtenu n'est pas aussi précis que le squelette obtenu par l'approche du squelettiseur.

Quelle que soit la méthode choisie, le squelette obtenu nécessite des traitements ultérieurs (polygonisation, amélioration de la précision) pour pouvoir conserver l'information significative.

V.3. Polygonisation.

Après avoir extrait les bords de parcelle, l'étape suivante est la polygonisation du squelette qui permet d'extraire et relier tous les segments de notre planche, ceci afin de s'approcher au mieux des traits dessinés par l'agent administratifs. La polygonisation est basée sur une approximation du polygone au moyen des moindres carrés **[8]**. L'élément central de cette méthode et son intérêt principal sont présentés ci-dessous. Si nous considérons une séquence de points à polygoniser, pour chaque point nouveau, nous réalisons un test entre deux segments (D1) et (D2). Le premier (D1) est calculé à partir de la méthode des moindres carrés appliquée à tous les points depuis le dernier changement de modèle. Le second (D2) est estimé par la méthode des moindres carrés caculée sur les N derniers points. N est déterminé comme coefficient de filtrage. Si l'angle entre (D1) et (D2) est supérieur à α, un changement de modèle est alors détecté. α est le coefficient de polygonisation . N et α dépendent du degré de polygonisation choisi.

Cette méthode a été comparée aux méthodes classiques de polygonisation **[9]**. Un de ses avantages principaux réside principalement dans le temps de traitement car les calculs nécessaires à chaque itération sont relativement limités. De plus, cette méthode est particulièrement adaptée à nos contours de parcelle car le bruit des segments est partiellement filtré par le calcul des moindres carrés. L'avantage principal de cette méthode repose sur le fait que l'orientation de la ligne originale et les points de polygonisation fournis par cet algorithme n'appartiennent pas nécessairement à la séquence de points de départ. Nous présentons à la figure suivante (Figure 6) les résultats de notre méthode de polygonisation appliquée à une parcelle.

Figure 6: Résultats de la polygonisation

V.4. Recalage des contours.

Les segments des contours obtenus par la méthode précédente sont connus avec une précision insuffisante à cause de la méthode de polygonisation utilisée. La position de chaque nœud (intersection de deux segments du contour) est trop imprécise. Pour réaliser un recalage de chaque nœud, donc de chaque segment, ceux-ci sont mis en correspondance dans l'image initiale avant squelettisation (avec l'épaisseur correspondante). Si il n'y a pas de correspondance, ils sont progressivement déplacés pour atteindre une correspondance la meilleure possible. Ici aussi, nous avons raisonnés par émission d'hypothèses, les données concernant les segments ne devenant valides qu'après une vérification dans l'image initiale.

VI. CONCLUSION.

Les résultats que nous présentons dans cette communication font état d'outils bas niveaux de vectorisation. Ils sont validés sur un nombre significatifs de àplanches cadastrales provenant de différentes municipalités. Cette démarche nous a permis de vérifier la fiabilité de nos algorithmes par rapport à la variabilité des représentations des planches.

Nos développements actuels visent à mettre au point un outil de reconnaissance de caractères multi-orientation du cadastre.

D'autre part, nous travaillons également sur l'implantation de ces différents outils selon une stratégie générale d'interprétation de documents.

REMERCIEMENTS

Ces travaux sont les résultats d'une collaboration entre la société MS2i, la région Haute Normandie et l'Université de Rouen.

BIBLIOGRAPHIE

[1] Ejiri.M, Kakumoto.S, Miyatake.T, Shimada.S and Iwamura.I: "Automatic Recognition of Engineering Drawings and Maps".Image Analysis Application edited by Rangachar Kasturi and Mohan M.trivedi.(1990) pp 73-126.

[2] Kasturi.R and Alemany.J "Information Extraction from Images of Paper-Based Maps". IEEE trans. on software Engineering Vol 14 n°5.(1988) pp 671-675.

[3] Joseph. S. H, Pridmore P., "Knowledge-Directed Interpretation of Mechanical Engineering Drawings", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol 14, pp 928-940.

[4] "Vectorisation de cartes numérisées" - Lefrere L., Menu E., Ogier J.M., Olivier C., Lecourtier Y. - 13è colloque GRETSI, Juan-lès-Pins, 16-20 sept. 1991, pp 161-164

[5] S.Susuki and T.Yamada. MARIS: "Map Recognition Input System" IAPR Workshop on computer Vision proceedings, Tokyo (1988) pp 421-426.

[6] Volet P., "Analyse et synthèse d'images de textures structurées", Thèse de Doctorat, Ecole Polytechnique de Lausanne, Mai 1987.

[7] "Extraction of road network from digitized maps", Ogier J.M., Olivier C., Lecourtier Y., 6 th EUropean SIgnal Processing COnference, EUSIPCO 92 August 24-27, 1992.

[8] Mullot R, "Segmentation d'images et extraction de primitives pour la reconnaissance optique de texte", Thèse de doctorat de l'université de Rouen, Janvier 1991.

[9] Wall K. and Daniellson P., "A fast sequential method for polygonal approximation of digitized curves", Computer Vision, Graphic and Image Processing, vol 28, N° 2, Nov 1984, pp 220-227.

[10] Taconnet B, Zahour A, Zhang S, Faure A, "Deux algorithmes de squelettisation", Actes du colloque RAE, Le Havre, BIGRE N° 68, pp. 68-76, 1990.