Traitement préliminaire de l'image d'un texte manuscrit en vue de sa reconnaissance: une méthode de sur-segmentation

Gilbert Maître, Stéphane Brunet*
IDIAP, c.p. 592, CH-1920 Martigny

Gianni Pante
GPIL, CH-1920 Martigny

Résumé: Dans le cas d'une écriture liée, la segmentation en caractères de l'image numérique d'un manuscrit, en tant que traitement préliminaire à la reconnaissance du texte, pose problème. Cet article décrit une méthode de sur-segmentation par rapport à la segmentation en caractères dont le but est de fournir, sous la forme d'une suite, un ensemble restreint mais suffisant de positions potentielles de caractères qui serviront d'hypothèses à l'étape de reconnaissance.

Supposant que les lignes du texte sont approximativement horizontales, la méthode procède à une segmentation en traits verticaux et horizontaux à partir du contour extraits de l'image. Les segments verticaux sont ensuite ordonnés pour fournir la suite des positions potentielles de caractères, à partir de laquelle il s'agira ensuite de reconnaître le texte en appliquant localement une méthode de reconnaissance des caractères. Le fonctionnement de la méthode est illustré sur quelques exemples de manuscrits anciens.

Mots clés: reconnaissance de l'écrit, segmentation de l'image d'un texte, manuscrits anciens

1 Introduction

La démarche classique pour reconnaître un texte écrit, représenté par une image numérique,comprend les deux étapes consécutives suivantes: (i) segmenter l'image en régions correspondant chacune à l'empreinte ("pattern") d'un caractère, (ii) reconnaître les caractères représentés par les segments d'image. Cette approche s'est avérée fructueuse dans des cas qui bénéficient de conditions particulières au niveau du texte proprement dit et/ou au niveau de la production du document écrit. Le succès est assuré par l'utilisation de méthodes de segmentation qui, pour isoler les caractères, profitent de ces conditions particulières et de la connaissance *a priori* y relative.

Ainsi, par exemple, la méthode de segmentation des codes postaux manuscrits proposée dans [CS90] profite du fait (et de la connaissance y relative) que les codes postaux traités ont un nombre de chiffres fixe. Il s'agit là d'une particularité au niveau du texte à reconnaître. Les cas de traitements où la segmentation profite de circonstances propices au niveau de la production du document sont nombreux. Parmi ces documents aux conditions favorables, nous citerons en premier lieu, les documents spécialement prévus pour l'OCR tels que les formulaires pourvus de cases contraignant la position des caractères manuscrits. Il y a aussi les documents écrits à la machine où les caractères sont frappés à intervalle régulier [Hen68]. Finalement, beaucoup de documents imprimés et certains documents manuscrits sont tels que les empreintes de caractères sont relativement bien séparées par un espace, dans lequel il y a parfois place pour une ligne qui, dans certains cas, peut être mise verticalement.

Dans ce dernier cas de figure, les méthodes de segmentation telles que projection verticale et extraction des composantes connexes ont fait leurs preuves [CL95]. Ces méthodes de base sont rendues robustes vis-à-vis de l'absence occasionnelle de séparation entre empreintes de caractères par l'ajout de pré-traitements [TA92], mais surtout de post-traitements. Ceux-ci interviennent soit au niveau de la segmentation (évaluation si chaque segment est un caractère potentiel, essentiellement sur des critères de taille), soit au niveau du texte reconnu (par exemple, correction lexicale [Cas86]).

Dans les cas où les caractères sont fréquemment liés, la segmentation est plus problématique. Une partie de la recherche en reconnaissance de l'écrit, s'appuyant sur les théories de la perception visuelle et du Gestalt [Lec93], poursuit dans l'approche classique en développant des méthodes de segmentation appropriées. Une autre partie, faisant référence au paradoxe: "il est nécessaire de segmenter pour reconnaître, mais il est aussi nécessaire de reconnaître pour segmenter" [Lec93], envisage des solutions alternatives [CL95], qui sont:

1. "segmentation puis reconnaissance d'entités

*stage de DESS ISA, Univ. de Rennes I, mars-juin 1995

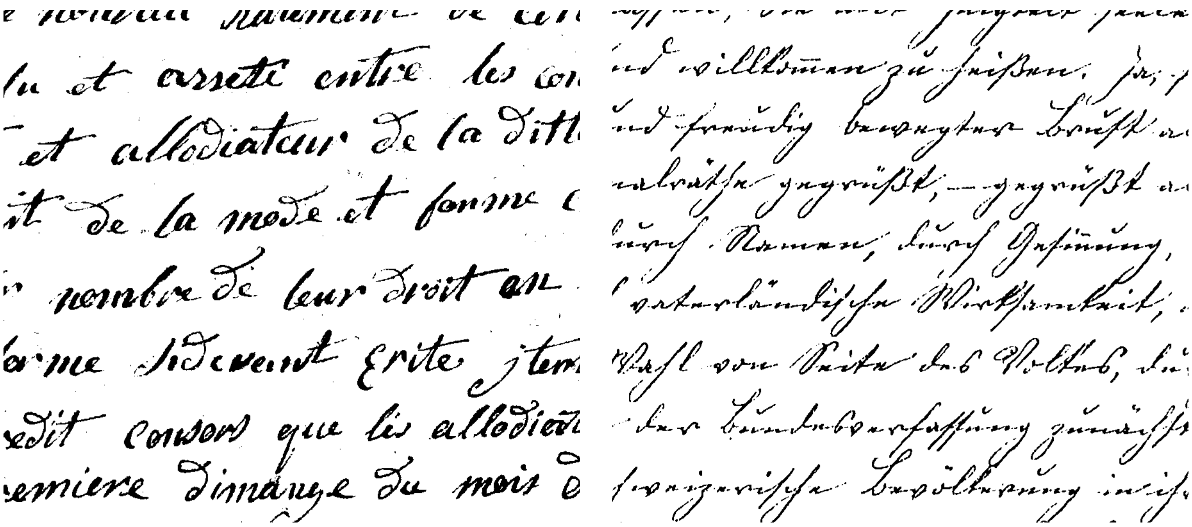

Figure 1: Images originales. A gauche, TEXTE1. A droite, TEXTE2.

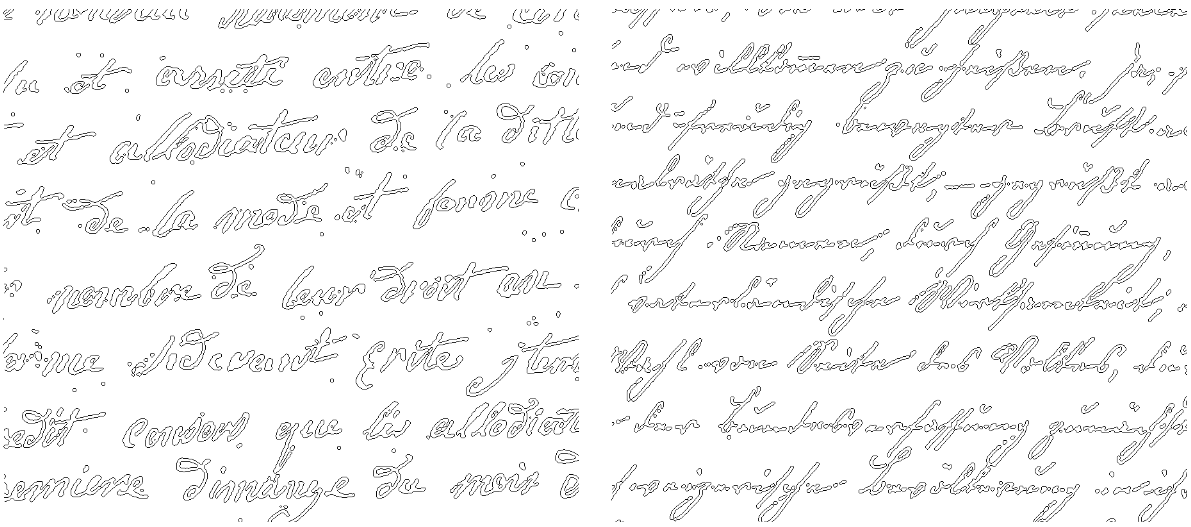

Figure 2: Images du contour. A gauche, TEXTE1. A droite, TEXTE2.

autres que les caractères". Certains travaux utilisent des entités appelées graphèmes ou pseudo-caractères, qui sont généralement des entités de taille inférieure au caractère. Une telle approche a été choisie déjà par Sayre en 1973 [Say73]. D'autres travaux utilisent les mots comme unités segmentales, c'était le cas déjà de Earnest en 1962 [Ear62].

2. "segmentation et reconnaissance simultanée des caractères". Les deux opérations sont réalisées de manière simultanée. La segmentation est un produit secondaire. Les travaux de Kovalevsky [Kov68] et ceux de Casey et Nagy [CN82] ont suivi une telle approche.

3. "sur-segmentation". La segmentation est réalisée en entités de taille plus petite que les caractères. Ces entités ne sont pas directement reconnues (comme dans l'approche 1), elles sont regroupées afin de générer des hypothèses pour la reconnaissance des caractères. Par rapport à l'approche 2, elle évite de faire une recherche "à l'aveugle" de l'image. Ce principe de sur-segmentation a été appliqué en partie dans des travaux tels que celui de Ameur *et al.* [ARPOL93]. Il a été utilisé de manière systématique dans les travaux de Schenkel *et al.*, dans leur méthode INSEG [SWG+93] et ceux de Ha *et al.* dans la partie de leur système consacrée au traitement local [HNB95].

C'est cette dernière alternative que nous avons choisie.

Dans la suite de cet article, nous décrivons les étapes de traitement constituant la méthode de sur-segmentation que nous proposons. Ce sont, dans l'ordre: extraction du contour (section 2), segmentation en traits verticaux et horizontaux (section 3) et, finalement, ordonnancement des traits verticaux (section 4). Nous terminons par une partie expérimentale concernant des documents manuscrits anciens (section 5). Des extraits d'images de deux documents servent d'illustration aux étapes de traitement. Ces extraits sont représentés à la figure 1. Le premier de ces documents, que nous désignons par TEXTE1, est une règle d'alpage en français, datant de 1748 (provenance: privée). Le deuxième, TEXTE2, est tiré du Protocole du Conseil National Suisse, en allemand, séance du 6 novembre 1848 (provenance: Archives Fédérales Suisses).

2 Extraction du contour

Partant d'une image binaire, la première étape de traitement réalise une extraction de la position et de l'orientation du contour de la figure ("foreground") de l'image.

L'approche poursuivie ici, qui est celle utilisée par Thomas Breuel dans son système de reconnaissance de caractères [Bre93], se démarque de celle qui est souvent appliquée et qui consiste à squelettiser la figure de l'image afin d'extraire le trait de l'écrit. Notre choix, qui rejoint l'opinion d'autres auteurs [TZF92], est motivé d'une part par l'expérience visuelle que la squelettisation peut parfois engendrer une déformation importante et, d'autre part, par les expériences faites sur la base "NIST Special Database 3" [GW92] avec le système de reconnaissance de caractères [Bre93]. Ces dernières ont en effet montré que l'utilisation d'opérations morphologiques pour normaliser l'épaisseur du trait (la squelettisation en est un cas particulier) amoindrit les scores de reconnaissance.

La position du contour de la figure de l'image pourrait être extraite simplement par une opération morphologique. Néanmoins, comme d'une part nous nous intéressons aussi à l'orientation du contour et que, d'autre part, il est intéressant d'avoir un traitement aussi applicable aux images en niveaux de gris, nous procédons de la manière suivante. D'abord, l'image est lissée par convolution avec une gaussienne. Puis, le gradient et le laplacien sont calculés. Enfin, les passages à zéro du laplacien pour lesquels le gradient est non nul sont déterminés. Ces points définissent la position du contour, et la direction du gradient en ces points définit son orientation. La figure 2 représente la position du contour extrait deso images originales de la figure 1 (avec une valeur d'écart-type de 2.0 pixels pour la gaussienne servant au lissage).

3 Segmentation en traits verticaux et horizontaux

La deuxième étape de traitement segmente le contour en parties verticales et horizontales, que nous appelons traits. Son but est de réaliser une sur-segmentation par rapport à la segmentation de l'image suivant les caractères.

Nous proposons deux versions de cette segmentation. La première, réduite à l'essentiel, démontre le principe. La deuxième, qui inclut deux éléments supplémentaires par rapport à la première, devrait s'avérer plus performante dans la pratique. Nous commençons par décrire la version simple.

3.1 Version élémentaire

De manière formelle, la segmentation résultant de cette étape de traitement peut être décrite comme

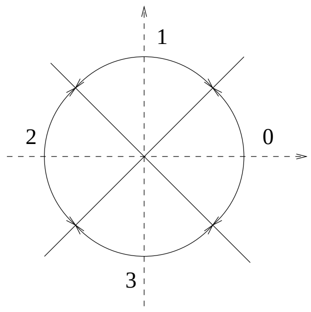

Figure 3: Domaines d'orientation simples.

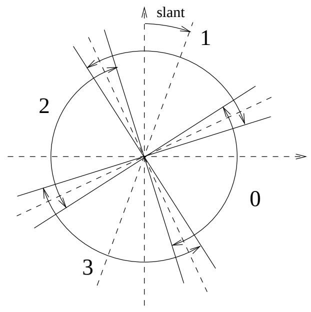

Figure 5: Domaines d'orientation évolués.

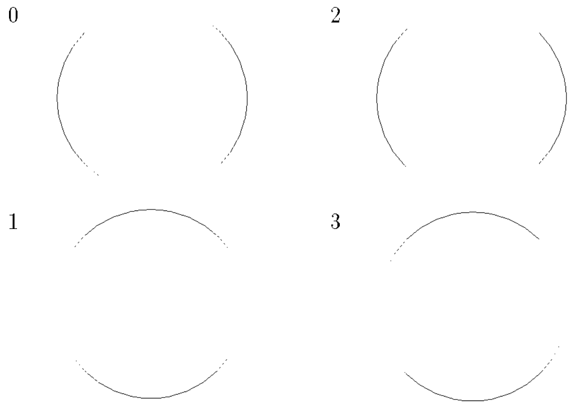

Figure 4: Segmentation simple du contour (intérieur et extérieur) d'un cercle.

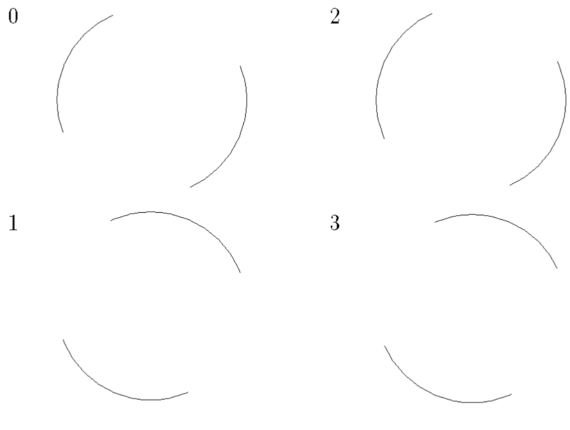

Figure 6: Segmentation avec hystérésis du contour (extérieur et intérieur) d'un cercle.

suit. Etant donné la partition de l'espace des valeurs d'orientation en quatre domaines de même taille centrés sur les directions (sens positifs et négatifs) des axes de coordonnées de l'image (voir figure 3), un segment est constitué de l'ensemble des points connexes du contour pour lesquels l'orientation du contour appartient au même domaine.

Pour aboutir à cette segmentation, nous utilisons l'algorithme suivant. D'abord, quatre images binaires sont créées, une par domaine d'orientation. Puis, pour chaque point contour, le domaine auquel appartient l'orientation du contour est calculé et la position du point est marquée dans l'image correspondant au domaine d'orientation. A ce stade du traitement, chaque image binaire représente donc la portion du contour dont l'orientation se trouve à l'intérieur du domaine associé à l'image. En dernière étape, les segments sont extraits, image par image, en utilisant un algorithme tel que l'algorithme de suivi de ligne [HS92] p.560.

Il y a donc quatre classes de segments correspondant aux quatre domaines d'orientation. Nous appelons segments verticaux droits les segments qui correspondent au domaine d'orientation qui est centré sur l'axe positif d'une ligne de l'image et segments verticaux gauches ceux du domaine opposé.

Cette méthode simple possède deux inconvénients. D'une part, les domaines d'orientation ne sont pas alignés sur l'inclinaison de l'écriture ("slant"), mais sur la direction verticale. D'autre part, le seuil entre domaines d'orientation provoque, en raison du bruit, l'apparition de petits segments parasites. La figure 4 démontre le phénomène dans le cas d'un cercle.

3.2 Version affinée

La modification par rapport à la version élémentaire réside dans l'utilisation de domaines d'orientation qui, d'une part, sont alignés sur l'inclinaison de l'écriture et, d'autre part, se superposent partiellement pour réaliser un mécanisme d'hystérésis et éliminer ainsi les segments parasites (voir figure 5).

L'hystérésis est mise en oeuvre en appliquant la règle suivante pour classifier de manière univoque les points du contour suivant les quatres

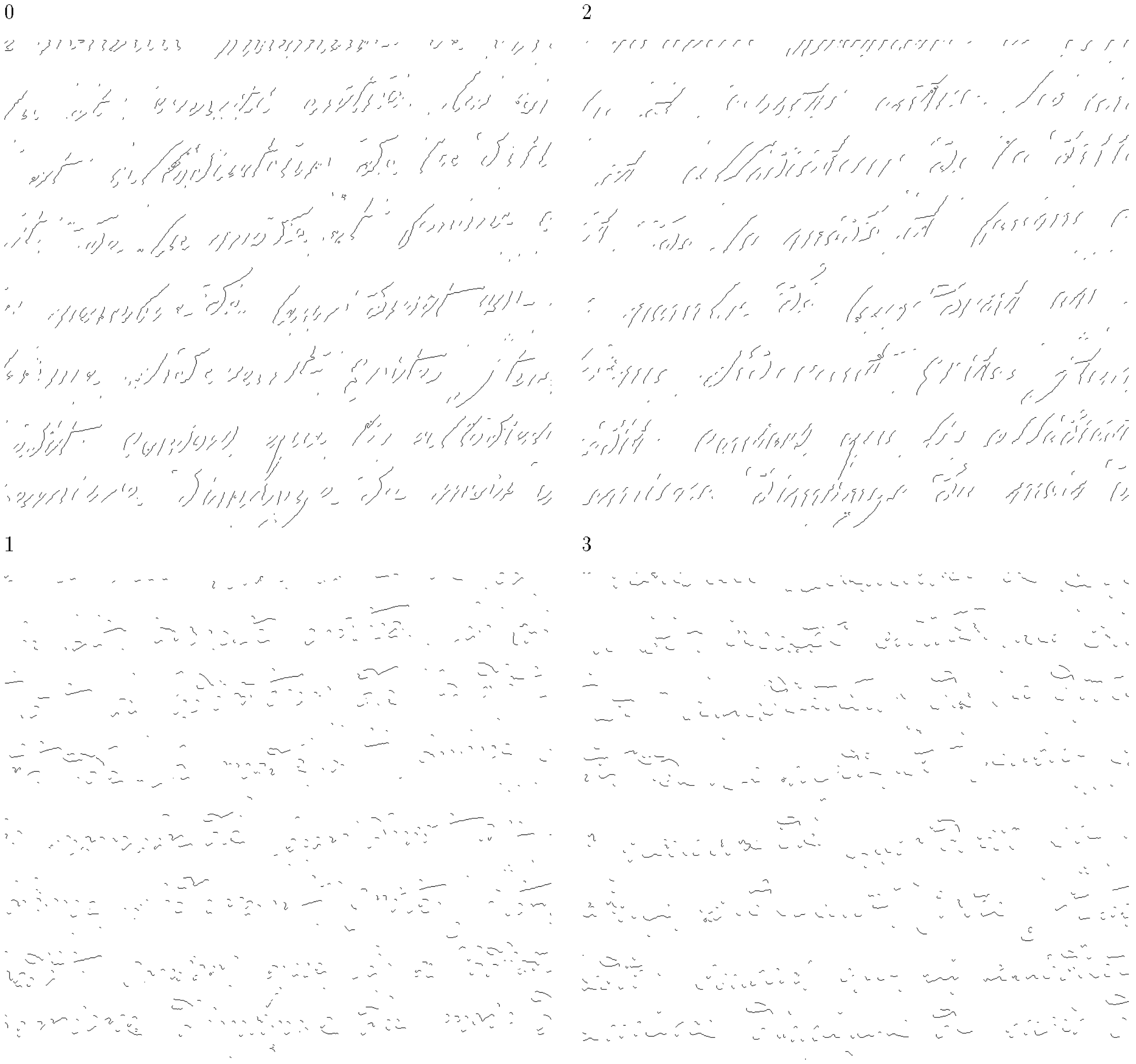

Figure 7: Images des segments de TEXTE1.

domaines d'orientation: aux points où l'orientation du contour appartient à deux domaines, on associe le même domaine d'orientation que celui du point qui est le plus proche le long du contour et pour lequel le domaine est univoque.

L'alignement des domaines d'orientation sur l'inclinaison de l'écriture requiert une estimée de cette grandeur. Différentes méthodes peuvent être utilisées pour fournir une telle estimée, par exemple, celle de Bozinovic et Srihari [BS89]. Nous mettons au point une méthode basée sur l'histogramme des orientations du contour.

La comparaison des figures 4 et 6 montrent le bénéfice de l'hystérésis sur la segmentation du contour d'un cercle. Les résultats obtenus en segmentant le contour de TEXTE1 et de TEXTE2 (figure 2) sont représentés dans les figures 7 et 8, respectivement.

4 Ordonnancement des traits verticaux

Le but de ce traitement est d'obtenir à partir des traits verticaux (gauches ou droits) une liste de positions ordonnées par ligne du texte, sur lesquelles on appliquera plus tard une méthode de reconnaissance des caractères.

Pour construire cette liste, nous utilisons une relation de voisinage entre traits verticaux, définie comme suit: deux traits verticaux sont voisins si et seulement si (i) il existe des points de l'un et l'autre traits se trouvant sur la même ligne d'image et (ii) tous ces points sont voisins sur

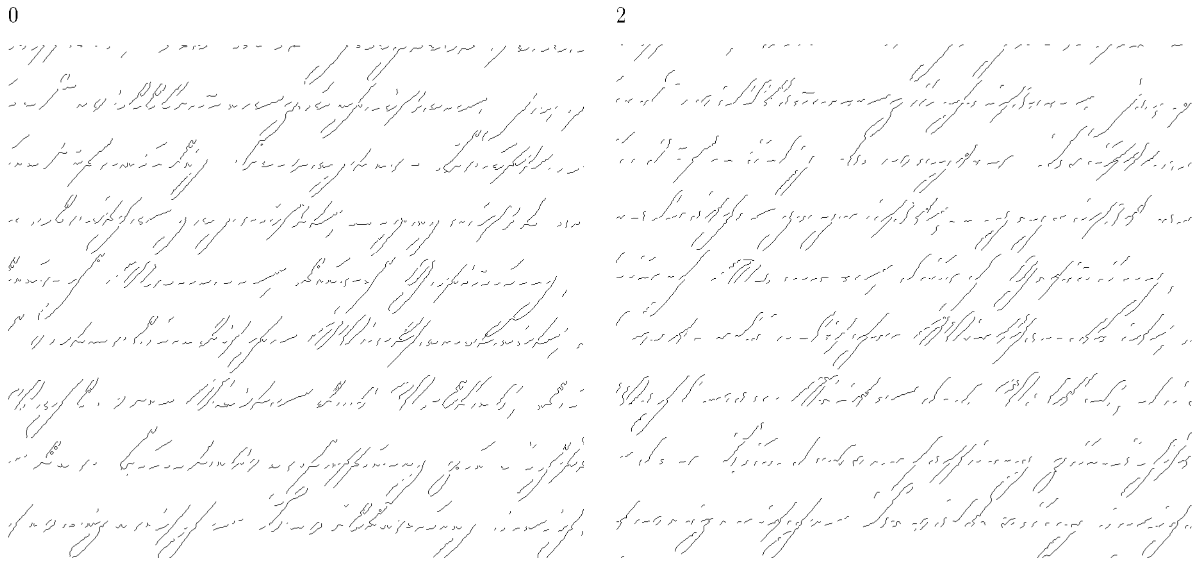

Figure 8: Images des segments de TEXTE2 (traits verticaux uniquement).

la ligne d'image, autrement dit, aucune paire de points n'est séparée par des points d'un autre segment. La figure 9 représente cette relation de voisinage entre traits verticaux gauches de TEXTE1.

La relation de voisinage que nous venons de définir permet de construire un graphe dans lequel chaque noeud représente un trait et chaque arc une relation de voisinage entre traits. Les connexions des arcs aux noeuds sont ordonnées suivant les lignes de l'image. En assimilant les bords gauche et droit de l'image à des traits verticaux, on obtient pour chacun d'eux une liste ordonnée de relations de voisinage. Ces listes correspondent approximativement aux lignes de texte.

Le graphe ainsi obtenu fournit la liste de positions, ordonnées ligne par ligne, sur laquelle nous allons appliquer localement une reconnaissance de caractères. Dans ce graphe, des traits appartenant à différentes lignes de texte sont parfois connectés et une ligne d'image est divisée parfois en plusieurs chemins. Néanmoins, nous ne voulons pas appliquer un post-traitement à ce graphe, visant à éliminer certaines connexions entre lignes et/ou à fusionner des chemins appartenant à la même ligne de texte, car ce post-traitement risque d'introduire des erreurs qui ne pourraient être corrigées que difficilement par la suite.

5 Expérimentation

Nous avons expérimenté cette méthode de segmentation sur des images de manuscrits anciens.

Notre intérêt pour ce genre de documents est motivé par deux aspects. Premièrement, les archives publiques offrent une grande variété de documents intéressants pour la recherche, à cause de leur diversité dans: la qualité du texte (orthographe,...), la régularité et le style de l'écriture, l'état de conservation du document. Deuxièmement, le monde technique moderne se doit de donner aux historiens, mais aussi au public, des outils facilitant l'accès à l'énorme héritage transmis par l'écriture.

Dans un premier temps, nous nous intéressons à des documents de bonne qualité caractérisés par une grande régularité de l'écriture. Pour notre recherche, nous avons commencé à constituer une base de données avec des documents provenant des Archives Cantonales Valaisannes et des Archives Fédérales Suisses. Les document originaux se trouvent sous forme de papier ou de microfilm. Pour l'instant, nous avons à disposition sous forme numérique une vingtaine de pages, écrites par deux personnes au total.

Comme il n'est pas possible d'évaluer quantitativement le résultat de la segmentation, nous avons appliqué nos algorithmes à quelques portions de pages pour en observer le comportement. Le résultat de la segmentation nous a paru satisfaisant dans le sens qu'il réalisait bien une sursegmentation par rapport à la segmentation en caractères.

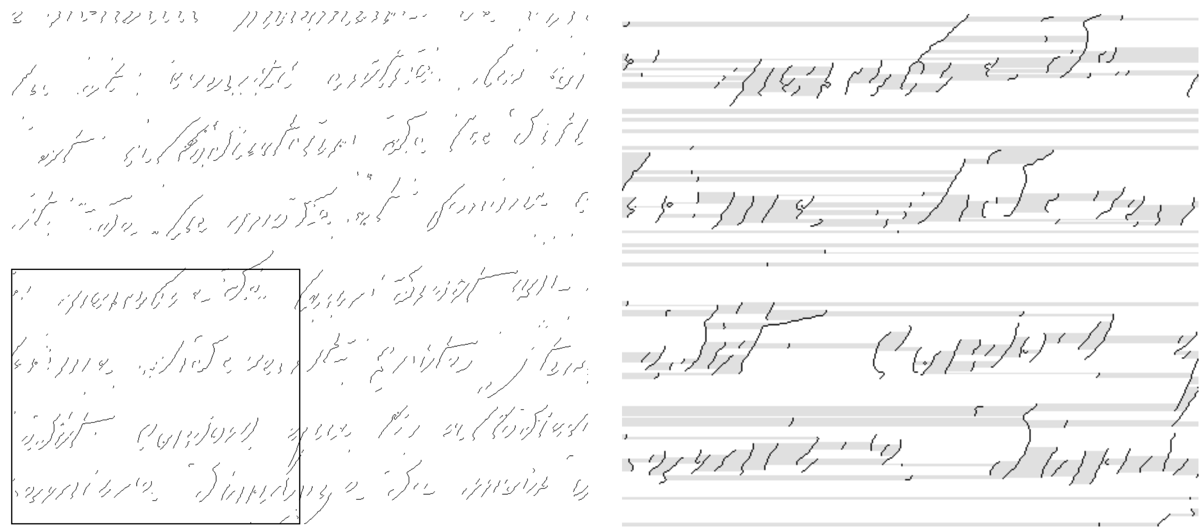

Figure 9: Relation de voisinage entre traits verticaux. A gauche: selection dans l'image des traits verticaux gauches de TEXTE1. A droite: relations de voisinage pour la sélection (représentées en gris).

6 Conclusion et perspectives

Dans cet article, nous avons présenté une méthode de sur-segmentation devant servir de traitement préliminaire pour la reconnaissance d'un texte manuscrit à partir d'une image numérique. La méthode est constituée d'une suite de traitements simples: extraction du contour, segmentation en traits verticaux et horizontaux, et ordonnancement des traits verticaux. Elle doit être considérée comme la première étape d'une approche ascendante (de local à global) de la reconnaissance de l'écrit.

La valeur de l'approche ne pourra être démontrée que lorsque nous aurons développés les traitements réalisant l'étape suivante, dont la fonction est la reconnaissance des caractères. Pour réaliser cette reconnaissance, nous envisageons, plutôt que de regrouper les segments, de travailler sur l'image du contour avec des méthodes robustes de mise en correspondence, telles que celles qui ont été étudiées et développées dans [Bre93, Mai95].

Remerciements

Le travail rapporté ici a été financé par le Fonds National Suisse pour la Recherche Scientifique, dans le cadre du projet no. 21-39576.93. Ce projet a été initié et dirigé par Thomas M. Breuel.

Les Archives Fédérales Suisses, en la personne de M. Comment, et les Archives Cantonales Valaisannes, par M. Truffer, ont mis à notre disposition les manuscrits qui ont servi à l'expérimentation. La numérisation des microfilms des Archives Fédérales a été assurée par la maison Quorum, Aldershot, Grande-Bretagne.

Laurence Likforman-Sulem nous a rendu attentifs à certaines références de littérature qui ont été utiles à notre travail et à l'élaboration de cet article. L'article dans sa forme actuelle a bénéficié des commentaires de Gérard Chollet, Eric Lecolinet, et Laurence Likforman-Sulem.

Nous adressons à toutes les personnes qui ont directement ou indirectement collaboré à ce travail nos sincères remerciements.

Références

[ARPOL93] A. Ameur, K. Romeo-Pakker, C. Olivier, and Y. Lecourtier. Segmentation et reconnaissance de textes manuscrits arabes. [ICO93], pages 80–82.

[Bre93] Thomas M. Breuel. Recognition of handprinted digits using optimal bounded error matching. [ICD93], pages 493–496.

[BS89] R.M. Bozinovic and S.N. Srihari. Off-line cursive word recognition. *IEEE Trans. Pattern Analysis and Machine Intelligence*, PAMI-11(1):68–83, January 1989.

[Cas86] R.G. Casey. Text OCR by solving a cryptogram. In *Proc. 8th ICPR*, pages 349–351, Oct. 1986.

[CL95] Richard G. Casey and Eric Lecolinet. Strategies in character segmentation: a survey. [ICD95], pages 1028–1033.

[CN82] R.G. Casey and G. Nagy. Recursive segmentation and classification of composite patterns. In *Proc. 6th ICPR*, pages 1023–, 1982.

[CS90] M. Cesar and R. Shinghal. Algorithm for segmenting handwritten postal codes. *Int. Journ. Man-Mach. Stud.*, 33(1):63, July 1990.

[Ear62] L.D. Earnest. Machine recognition of cursive writing. In C. Cherry, editor, *Information Processing*. Butterworth, London, 1962.

[GW92] M. D. Garris and R. A. Wilkinson. NIST Special Database 3: Binary images of handwritten segmented characters, 5.25" CD-ROM. available from: Standard Reference Data, National Institute of Standards and Technology, 221/A323, Gaithersburg, MD 20889, 1992.

[Hen68] R.B. Hennis. The IBM 1975 optical page reader: system design. *IBM Journ. Res. & Dev.*, pages 346–353, Sept. 1968.

[HNB95] Thien M. Ha, Dieter Niggeler, and H. Bunke. A system for segmenting and recognising totally unconstrained handwritten numeral strings. [ICD95], pages 1003–1009.

[HS92] R. M. Haralick and L. G. Shapiro. *Computer and robot vision*, volume 1. Addison-Wesley, 1992.

[ICD93] *Second International Conference on Document Analysis and Retrieval (ICDAR'93)*, 1993.

[ICD95] *Third International Conference on Document Analysis and Retrieval (ICDAR'95)*, 1995.

[ICO93] *Sixth International Conference on Handwritting and Drawing (ICOHD'93)*, 1993.

[Kov68] V.A. Kovalevsky. *Character readers and pattern recognition*. Spartan Books, Washington D.C., 1968.

[Lec93] Eric Lecolinet. Top-down and bottom-up strategies for cursive script recognition. [ICO93], pages 117–119.

[Mai95] Gilbert Maitre. Experiments with robust similarity measures for OCR. Technical Report 95-03, IDIAP, June 1995.

[Say73] K.M. Sayre. Machine recognition of handwritten words: a project report. *Pattern Recognition*, 5:213–228, 1973.

[SWG+93] Markus Schenkel, Haim Weissmann, Isabelle Guyon, Craig Nohl, and Don Henderson. Recognition-based segmentation of on-line handprinted words. In *Advances in Neural Information Processing Systems 5*, pages 723–730. Morgan Kaufmann, Denver, 1993.

[TA92] S. Tsujimoto and H. Asada. Major components of a complete text reading system. *Proc. IEEE*, 80(7):1133–1149, July 1992.

[TZF92] B. Taconnet, A. Zahour, and A. Faure. A new global off-line recognition method for handwritten words. In S. Impedovo and J.C. Simon, editors, *From Pixels to Features III: Frontiers in Handwriting Recognition*, pages 327–338. North Holland, 1992.