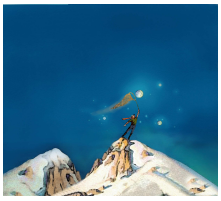


printemps des sciences
Avec le soutien de la Ministre de l'Enseignement supérieur et de la Recherche scientifique

Sciences extrêmes 19 - 25 mars 2007

ULB

RECONNAISSANCE OPTIQUE DE CARACTÈRES

Dirk Amadori
Département d'Informatique

La reconnaissance optique de caractères est un procédé informatique permettant la conversion de documents papier en texte éditable sur un ordinateur.

Il existe 2 types de reconnaissance :
1) Reconnaissance hors-ligne : on travaille sur des caractères déjà écrits ou imprimés.
On possède alors une représentation statique des caractères.

2) Reconnaissance en-ligne : on travaille sur des caractères en train d'être écrits.
En plus de la représentation du caractère, on dispose d'informations temporelles, telles que la vitesse d'écriture et l'ordre dans lequel les traits composant le caractère ont été tracés.

Applications possibles :
- Tri postal automatisé selon le code postal.
- Encodage de tout type de documents (factures, chèques, virements, ...).
- Remplacer le clavier par un périphérique tactile permettant la reconnaissance en-ligne pour les PDA.

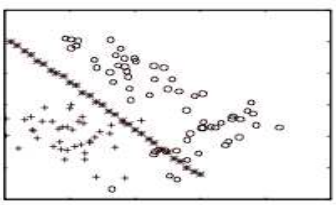
Un tel système est composé de 4 grandes étapes:

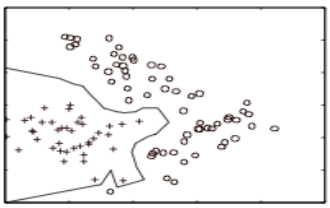
1) Pré-traitement
- nettoyage de l'image : suppression des pixels parasites.
- segmentation du document en lignes, et de chaque ligne en caractères.
- normalisation des caractères.

Segmentation d'un mot en caractères

2) Extraction de caractéristiques
Une caractéristique est une valeur représentant le caractère.
L'extraction de caractéristiques est en quelque sorte un résumé du caractère.
Il existe de nombreuses caractéristiques possibles, toutes n'étant pas aussi bonnes (par exemple le rapport hauteur sur largeur ne permet pas de décrire de manière assez précise un caractère).
Les caractéristiques du caractère sont ensuite regroupées dans un vecteur de caractéristiques pouvant être représenté dans un espace à d-dimensions (d est le nombre de caractéristiques retenues).

3) Classification
La classification se charge de découper en classes l'espace à d-dimensions contenant les vecteurs de caractéristiques. A chaque classe correspond une lettre (ou tout autre symbole).





4) Post-traitement
Principalement vérification orthographique des mots à l'aide d'un dictionnaire, ainsi qu'une vérification grammaticale, ...