Reconnaissance optique de caractères

La **reconnaissance optique de caractères** (ROC), en anglais *optical character recognition* (**OCR**), encore appelée vidéocodage (traitement postal, chèque bancaire) ou *océrisation*, désigne les procédés informatiques pour la traduction d'images de textes imprimés ou dactylographiés en fichiers de texte.

Un ordinateur réclame pour l'exécution de cette tâche un logiciel d'OCR. Celui-ci permet de récupérer le texte dans l'image d'un texte imprimé et de le sauvegarder dans un fichier pouvant être exploité dans un traitement de texte pour enrichissement, et stocké dans une base de données ou sur un autre support exploitable par un système informatique.

Sommaire

Histoire

La première machine d'OCR fut créée par Gustav Tauschek, un ingénieur allemand, en 1929. Elle contenait un détecteur photosensible qui pointait une lumière sur un mot lorsqu'il correspondait à un gabarit contenu dans sa mémoire.

En 1950, Frank Rowlett, qui avait cassé le code diplomatique japonais PURPLE, demanda à David Shepard, un cryptanalyste de l'AFSA (prédécesseur de la NSA américaine), de travailler avec Louis Tordella pour faire à l'agence des propositions de procédures d'automatisation des données. La question incluait le problème de la conversion de messages imprimés en langage machine pour le traitement informatique. Shepard décida qu'il devait être possible de construire une machine pour le faire, et, avec l'aide de Harvey Cook, un ami, construisit « Gismo » dans son grenier pendant ses soirées et ses week-ends. Le fait fut rapporté dans le *Washington Daily News* du 27 avril 1951 et dans le *New York Times* du 26 décembre 1953 après le dépôt du brevet numéro 2 663 758. Shepard fonda alors Intelligent Machines Research Corporation (IMR), qui livra les premiers systèmes d'OCR au monde exploités par des sociétés privées. Le premier système privé fut installé au Reader's Digest en 1955, et, de nombreuses années plus tard, fut offert par le Readers Digest au Smithsonian, où il fut mis en exposition. Les autres systèmes vendus par IMR à la fin des années 1950 comprenaient un lecteur de bordereau de facturation à l'Ohio Bell Telephone Company et un numériseur (scanner de documents) à l'US Air Force pour la lecture et la transmission par télex de messages dactylographiés. IBM et d'autres utilisèrent plus tard les brevets de Shepard.

Depuis 1965, la Poste des États-Unis utilise pour trier le courrier des machines OCR dont le principe de fonctionnement a été imaginé par Jacob Rabinow, un inventeur prolifique. La Poste canadienne utilise des systèmes OCR depuis 1971. Les systèmes OCR lisent le nom et l'adresse du destinataire au premier centre de tri automatisé, et impriment sur l'enveloppe un code-barres fondé sur le code postal. Les lettres n'ont plus qu'à être triées dans les centres suivants par des trieuses moins coûteuses qui n'ont qu'à lire le code-barres. Pour éviter toute interférence avec l'adresse lisible qui peut se trouver n'importe où sur la lettre, une encre spéciale est utilisée, qui est clairement visible sous une lumière UV. Cette encre semble orange dans des conditions d'éclairage normales.

Il fallut attendre 1974 pour qu'un scientifique rassemble ces nouvelles connaissances dans une technologie qui permettrait aux aveugles de lire des documents enregistrés sur un support informatique. On sort du domaine précis de la reconnaissance optique de caractère pour l'appliquer en utilisant de nouvelles technologies. Pour cela, ce scientifique, du nom de Ray Kurzweil, créa un synthétiseur vocal pour « dire » le texte et améliora les procédés de numérisation. En 1976, le prototype fut fini, et pour l'anecdote, c'est Stevie Wonder qui finança le projet. Le début de la commercialisation de la « reading machine » eut lieu en 1978. Deux ans plus tard, Xerox acheta la société.

Apprentissage

Les premiers systèmes avaient besoin d'un « apprentissage » (la collecte d'échantillons connus pour chaque caractère) pour lire une police de caractères donnée. Mais aujourd'hui, il est courant de trouver des systèmes « intelligents » qui peuvent reconnaître la plupart des polices avec un haut niveau de précision [réf. nécessaire].

Fonctionnement

Un système OCR part de l'image numérique réalisée par un scanner optique d'une page (document imprimé, feuillet dactylographié, etc.), ou un appareil photo numérique, et produit en sortie un fichier texte en divers formats (texte simple, formats de traitements de texte, XML..., par exemple le format standardisé ALTO).

Certains logiciels tentent de conserver l'enrichissement du texte (corps, graisse et police) ainsi que la mise en page, voire de rebâtir les tableaux et d'extraire les images.

Certains logiciels comportent, en outre, une interface pour l'acquisition numérique de l'image.

Jusqu'à une date récente, le fonctionnement des systèmes OCR performants était peu connu car protégé par le secret industriel ; les logiciels open-source disponibles (ex : GOcr) étant plutôt l'œuvre d'amateurs. La publication en open-source de systèmes performants (en particulier Tesseract en 2006) a quelque peu changé cette situation.

Les étapes de traitement peuvent être schématisées ainsi :

1. **Préanalyse** de l'image : le but est d'améliorer éventuellement la qualité de l'image. Ceci peut inclure le redressement d'images inclinées ou déformées, des corrections de contraste, le passage en mode bicolore (noir et blanc, ou plutôt papier et encre), la détection de contours.
2. **Segmentation** en lignes et en caractères (ou Analyse de page) : vise à isoler dans l'image les lignes de texte et les caractères à l'intérieur des lignes. Cette phase peut aussi détecter le texte souligné, les cadres, les images.
3. **Reconnaissance** proprement dite des caractères : après normalisation (échelle, inclinaison), une instance à reconnaitre est comparée à une bibliothèque de formes connues, et on retient pour l'étape suivante la forme la plus « proche » (ou les N formes les plus proches), selon une distance ou une vraisemblance (*likelihood*). Les techniques de reconnaissance se classent en quelques grands types[1] :
    1. Classification par Caractéristiques (*Features*) : une forme à reconnaître est représentée par un vecteur de valeurs numériques - appelées *features* en anglais - calculées à partir de cette forme. Le nombre de *features* est de l'ordre de 100 à 300. Si les *features* sont bien choisies, une classe de caractères (par exemple l'ensemble des A majuscules) sera représentée par un « nuage » contigu de

points dans l'espace vectoriel des *features*. Le rôle du classificateur est de déterminer à quel nuage (donc à quelle classe de caractères) la forme à reconnaitre appartient le plus vraisemblablement. La classification fait généralement appel à divers types de réseaux de neurones artificiels entrainés sur de vastes bases de formes possibles.
   2. Méthodes métriques : consistent à comparer directement la forme à reconnaître, au moyen d'algorithmes de distance, avec un ensemble de modèles appris. Ce type de méthode est peu utilisé et peu valorisé par les chercheurs, car souvent plus naïf et vraisemblablement moins efficace que les méthodes à base de *features*.
   3. Méthodes statistiques : dans le domaine de la reconnaissance d'écriture manuscrite, il est fréquemment fait appel aux méthodes probabilistes/statistiques comme les chaînes de Markov.
4. **Post-traitement** utilisant des méthodes linguistiques et contextuelles pour réduire le nombre d'erreurs de reconnaissance : systèmes à base de règles, ou méthodes statistiques basées sur des dictionnaires de mots, de syllabes, de N-grammes (séquences de caractères ou de mots). Dans les systèmes industriels, des techniques spécialisées pour certaines zones de texte (noms, adresses postales) peuvent utiliser des bases de données pour éliminer les solutions incorrectes.
5. Génération du format de sortie, avec la mise en page pour les meilleurs systèmes.

Domaine de recherche

Un problème particulièrement ardu pour les ordinateurs et les humains est celui des anciens registres religieux des baptêmes et des mariages, qui contiennent surtout des noms, où les pages peuvent être endommagées par le temps, l'eau ou le feu, et les noms peuvent être obsolètes ou écrits selon d'anciennes graphies. Les techniques informatiques de traitement de l'image peuvent aider les humains dans la lecture de textes extrêmement difficiles, comme le palimpseste d'Archimède ou les manuscrits de Qumrân. Des approches coopératives où les ordinateurs assistent les humains et vice-versa constituent un domaine de recherche intéressant.

La reconnaissance de caractère est un domaine actif de recherche pour la science informatique depuis la fin des années 1950. Au début, on pensait qu'il s'agissait d'un problème facile, mais il apparut qu'il s'agissait d'un sujet beaucoup plus intéressant. Il faudra encore de nombreuses décennies aux ordinateurs, s'ils y parviennent un jour, pour lire tous les documents avec la même précision que les êtres humains.

Principaux logiciels de reconnaissance optique de caractères

Logiciels libres[2]

- GOCR (Unix, Windows)
- OCRopus (Unix)
- Tesseract (Unix, Windows)

Logiciels propriétaires

Logiciels freeware

- Moredata, logiciel freeware qui emploie tessnet (Windows).
- MoredataFast (Windows).

Payants

- Ad'doc IIM (Windows) reconnaissance dactylographié et manuscrit Www.imds.ca
- Adobe Acrobat Professional (Windows, Mac OS)
- ExactScan ExactScan Pro et OCRKit (Mac OS)
- ABBYY FineReader (Unix, Windows, Mac OS)
- Nuance Omnipage (Windows)
- I.R.I.S. Readiris (Unix, Windows, Mac OS)

- Nicomsoft OCR (Windows, Unix)
- Omnipage PRO (Windows 95 et suivants)

Notes et références

1. Principles of Pattern Classification: Statistical, Neural Net and Syntactic methods of getting robots to see and hear - Lecture Notes by Dr Michael D. Alder, University of Western Australia, 1994
2. libres, sous licences GNU GPL ou Apache.

Bibliographie

- *Reconnaissance de l'imprimé*, H 1348, par Philippe Lefèvre, éditions Techniques de l'Ingénieur.
- *Combining Multi-Scale Character Recognition and Linguistic Knowledge for Natural Scene Text OCR (http://liris.cnrs.fr/publis/?id=5603)*, Khaoula Elagouni [Orange Labs] , Christophe Garcia [LIRIS] , Franck Mamalet [Orange Labs] , Pascale Sébillot [IRISA]

Voir aussi

Articles connexes

- Reconnaissance intelligente de caractères
- Reconnaissance de l'écriture manuscrite
- Reconnaissance de formes
- Analyse discriminante
- Table des caractères Unicode - reconnaissance optique de caractères
- Base de données MNIST, un jeu de données standard pour la reconnaissance d'écriture manuscrite