

Une étude sur le choix des caractéristiques pour la représentation de caractères imprimés

Jean-Luc Henry

Equipe de Traitement d'Images, Vision et Parallélisme

Université des Antilles et de la Guyane

Département de Mathématiques et Informatique

97159 Pointe-à-Pitre, Guadeloupe

Jean-Luc.Henry@univ-ag.fr

Résumé

La représentation de caractères de façon fiable avant traitement (reconnaissance, apprentissage, classification, etc.) n'est pas totalement résolue. La difficulté subsiste dans le choix des critères pour caractériser une forme de façon non ambiguë. Aucune méthode aussi sophistiquée soit elle, ne peut à elle seule résoudre tous les problèmes. L'évolution de la recherche dans ce domaine s'oriente vers une tendance à la coopération de méthodes.

Cet article présente une étude sur le choix de la meilleure caractéristique pour représenter un caractère. Après avoir appliqué à un grand nombre de caractères nos algorithmes de recherche de contour, de génération de profils et de projections (partielles ou totales), nous arrivons à la conclusion qu'il est nécessaire de combiner plusieurs caractéristiques pour représenter un caractère.

Abstract

The representation of characters in a reliable way before processing (recognition, training, classification, etc.) is not completely resolved. The difficulty subsists in the choice of criteria to characterize a pattern in a non ambiguous way. No method as sophisticated as it, can not solve alone all problems. The evolution of the research in this field orients toward a tendency to the cooperation of methods. This paper presents a study on the choice of the best characteristic to represent a character. After having applied on a large number of characters our algorithms of contour research, of profiles and projection generation (partial or total), we arrive to the conclusion that it is necessary to combine several characteristics to represent a character.

1 Introduction

Dans cette étude, nous cherchons une description non ambiguë et la plus complète possible des formes de caractères. Nous avons exclu les primitives car elles nous paraissent moins robustes que les attributs et elles sont trop sensibles aux bruits et aux ruptures de tracés. Ce choix

s'avère prépondérant dans le cas où l'on envisage d'améliorer les critères de l'étape de reconnaissance du système. Les attributs ne nécessitent ni d'interprétations ni de phase de segmentation particulière ; nous allons les évaluer en nous appuyant sur une étude expérimentale. Nous avons limité notre étude aux caractéristiques issues du contour, des profils et des projections (partielles ou non). Nous fixons le même protocole pour tous les attributs afin de mesurer leur pouvoir discriminant. Nous utilisons pour chaque catégorie de tests un vecteur de caractéristiques de dimension 64.

La méthode de comparaison utilisée est celle des 5-ppv, car les k plus proches voisins sont des outils suffisamment éprouvés et robustes.

2 Recherche d'une description non ambiguë des formes de caractères

A partir d'une étude expérimentale, nous expliquons les choix des attributs utilisés. Nous considérons que les attributs sont plus robustes que les primitives. En effet, il n'y a pas d'erreurs dans l'extraction d'un attribut alors que la segmentation d'une primitive peut échouer et conduire à une fausse interprétation du caractère. Le choix des attributs par rapport aux primitives s'explique par l'objectif d'établir une coopération entre l'étape de reconnaissance et l'étape de correction contextuelle du système. En effet, il serait alors possible de modifier les décisions prises par l'étape de reconnaissance (à partir des attributs) en fonction des résultats fournis par l'étape d'analyse contextuelle (selon les échecs recensés). En outre, l'utilisation de primitives mal reconnues induirait de fausses corrections. En utilisant plutôt des attributs nous facilitons la recherche des causes d'erreurs dans la chaîne de traitement de l'information.

Notre expérimentation va être construite sur un échantillon de polices différentes (environ 200 fontes) sur plusieurs déclinaisons (gras, italique, inclinaison et homothéties différentes, ...) soit un total de 10700 caractères tous différents. Elles proviennent de pages tirées du livre de référence des peintres en lettres et des

