Tutoriel

Systèmes de reconnaissance de caractères pour les non-experts

par ***Nawwaf N. Kharma & Rabab K. Ward***
Université de Colombie-Britannique

This tutorial paper presents an overview of the field of character recognition by providing answers to the following questions:
- What does a character recognition system do?
- How does it do it i.e. what are its functional components?

The answers are meant to shed some light onto the field. Finally, what the authors believe are the two main open problems of character recognition are briefly described.

Cet article présente un aperçu des technologies de reconnaissance de caractères en vous fournissant les réponses aux questions suivantes:

- Quel est le rôle d'un système de reconnaissance de caractères?

- Quel en est le fonctionnement?

Les réponses vous aideront à comprendre cette nouvelle technologie. De plus, deux gros problèmes connus de la reconnaissance de caractères seront décrits brièvement.

1.0 Qu'est-ce qu'un système de reconnaissance de caractères?

La reconnaissance de caractères est le traitement machine (bi-dimensionnel) de formes d'entrée de texte en vue de produire certaines données de sortie compréhensibles. De fait, les systèmes de reconnaissance de caractères constituent un sous-ensemble des systèmes de reconnaissance des formes.

Les entrées peuvent provenir tant d'appareils en ligne que d'appareils hors ligne. Les appareils en ligne fonctionnent avec un stylo et comprennent des affichages-tablettes et des tablettes graphiques. Ces tablettes peuvent fournir l'ordre temporel des points qui constituent les lignes de texte. Certaines tablettes fournissent d'autres renseignements, notamment la vitesse (de l'écriture) et la pression (exercée par l'utilisateur). Par ailleurs, les appareils hors ligne comprennent des dispositifs de balayage de type à plat, manuel et papier. Ils rendent une image essentiellement sous forme de pixels en mode point.

Un système de reconnaissance de caractères (SRC) accepte les données de sortie provenant d'un équipement en ligne ou hors ligne comme des données d'entrée, en assure le traitement et produit des données de sortie compréhensibles. Parmi les formes de données de sortie possibles, mentionnons les séquences de symboles (p. ex. « O U I »), la date sur un chèque (p. ex. 14 février 1994) et la validité ou la non-validité d'une signature.

2.0 Quels sont les composants « fonctionnels » d'un système de reconnaissance de caractères?

Un SRC peut être effectivement morcelé en plusieurs composants. L'un de ces composants se charge des fonctions de « pré-traitement », comme la normalisation [1] et l'amincissement [2]. Une fois que la forme d'entrée a été prétraitée, un autre composant l'accepte et en extrait les attributs caractéristiques [3]. Les caractéristiques ainsi extraites sont utilisées par un composant de « classification » (comme un réseau neuronal [4]) pour attribuer une étiquette à la forme. Toutes les fonctions menées après la classification (initiale) font partie du « post-traitement ». Il est bon de noter :

- qu'il n'y a pas nécessairement de composants fonctionnels dans tous les SRC. Un SRC peut prendre charge de la fonction de classification sans avoir auparavant explicitement extrait les caractéristiques au moyen, par exemple, d'une certaine forme quelconque d'appariement par référence [5].
- que des composants fonctionnels ne sont pas toujours mis en place à titre de composants incompatibles. Ainsi, un objet logiciel peut extraire des caractéristiques et les classifier simultanément [6].
- que des composants fonctionnels n'interviennent pas nécessairement en séquence. De fait, dans de nombreuses applications, un nombre important d'extractions de caractéristiques sont effectuées avant la segmentation.

Quoiqu'il en soit, la grande majorité des SRC comprennent au moins trois des quatre composants décrits précédemment. Chacun des quatre composants fonctionnels d'un SRC est décrit avec plus de précision dans les paragraphes qui suivent.

2.1 Pré-traitement

Le pré-traitement inclut toutes les fonctions effectuées avant l'extraction des caractéristiques pour produire une version « nettoyée » de l'image d'origine afin qu'elle puisse être utilisée directement et efficacement par le composant d'extraction de caractéristiques du SRC. Ainsi, le pré-traitement comprend les étapes qui suivent :

A - Réduction du bruit (figure 1)

Le bruit, une erreur aléatoire dans la valeur de pixel, est une valeur découlant habituellement de la reproduction, de la numérisation et de la transmission de l'image originale. Le bruit peut être réparti en trois catégories : bruit dépendant du signal, bruit non dépendant du signal et bruit noir et blanc. Le bruit ne peut pas toujours être entièrement supprimé; on utilise souvent le lissage pour remplacer la valeur d'un pixel par la moyenne des valeurs des pixels entourant (et incluant) le pixel d'origine. Lorsqu'il s'agit d'images balayées, le lissage peut provoquer du maculage, et lorsqu'il est appliqué sur du texte en ligne, peut provoquer du découpage des points d'extremité.

Figure 1: Image d'une signature avant et après la réduction du bruit

B - Schématisation (figure 2)

Un texte est composé de lignes qui peuvent être d'un point d'épaisseur, comme c'est le cas pour la plupart des sources en ligne, dont les ordinateurs à stylet. Les images en ligne provenant des scanners ont cependant habituellement plusieurs points d'épaisseur. Les informations les plus pertinentes sur les lignes ne sont pas reliées à l'épaisseur de la ligne. Ainsi, l'amincissement des lignes en supprimant tous les pixels redondants jusqu'à ce que l'épaisseur ne soit plus que d'un point peut constituer une procédure très utile. Il faut alors se poser la question sui-

vante: quels sont les pixels redondants, et comment peut-on les supprimer de la ligne d'origine?

En règle générale, une procédure d'amincissement est évaluée selon sa capacité de contrôler des lignes de l'image d'origine sans, en même temps,

- fragmenter une ligne déjà continue en la divisant en plusieurs lignes isolées,
- découper les extrémités de la ligne centrale,
- introduire de nouvelles caractéristiques (p. ex. une courbe) qui n'étaient pas présentes à l'origine, ou
- supprimer ou remplacer une caractéristique (p. ex. en remplaçant une boucle par une ligne simple).

Un algorithme efficace d'amincissement est décrit en [7]. Cet algorithme enlève essentiellement des couches de pixels du contour de l'image d'origine de la ligne tout en évitant le découpage et la fragmentation des lignes.

Figure 2 : Le mot « Tanzanie » en arabe, après l'application d'un algorithme d'amincissement [8]

C- Normalisation

Les formes, soit les mots, peuvent prendre différentes formats, être placées à différents endroits (à l'intérieur d'une image), et sont souvent pivotées jusqu'à 180 degrés. Il est donc souvent nécessaire d'effectuer une opération de normalisation avant d'entamer toute extraction (ou appariement de formes). Les démarches de normalisation peuvent être réparties selon les groupes suivants :

- Techniques de moment invariant [9]
- Descripteurs de Fourier [10]
- Techniques reliées aux contours [1]
- Analyse vectorielle [11]

Ces démarches ont en commun les caractéristiques suivantes :

- Elles normalisent le format du caractère en divisant la caractéristique reliée au format utilisée, quelle qu'elle soit, par la longueur totale du caractère.
- Elles normalisent la position du caractère en déplaçant le centre des coordonnées vers un point qui se trouve à une position fixe sur le caractère ou proche du caractère, soit le point médian, ou vers le point d'origine du caractère en question.

La normalisation de l'orientation du caractère constitue toutefois une opération plus compliquée que les deux procédures précédentes, et elle est effectuée de maintes façons très différentes.

D - Segmentation (figure 3)

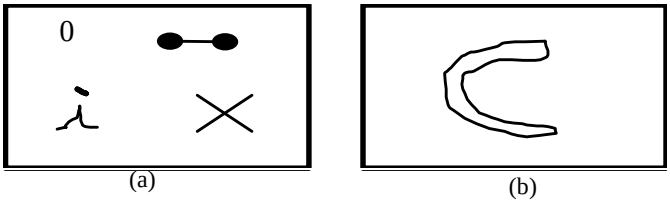
Les caractères peuvent être produits en lettres attachées. Ils peuvent également se chevaucher. Lorsqu'il est question d'un système de reconnaissance de caractères qui doit parvenir à identifier des caractères (au lieu d'identifier simplement des mots complets), il est nécessaire de déterminer (approximativement) où débute et où prend fin un caractère. C'est essentiellement le but de la segmentation. Il existe plusieurs méthodes pour contrer le problème de la segmentation :

- La pré-segmentation signifie (souvent) que des caractères sont déjà séparés les uns des autres. C'est ce qui se produit habituellement lorsque le texte est imprimé, ou lorsque l'auteur doit écrire les caractères dans des encadrés ou sans les réunir. (p. ex. avec le Palm III).
- La recherche de discontinuités : pour trouver des discontinuités entre les lettres ou, du moins, les lignes de raccordement. Pour avoir un aperçu de techniques axées sur la discontinuité, dont les matrices de caractères « stogrammes » [16], les longueurs de plage et les enveloppes convexes, reportez-vous à la note [13]. Toutes ces techniques fonctionnent par analyse des relations géométriques entre les divers composants du texte.
- Il existe des systèmes [14] qui permettent de classifier sans segmenter explicitement le mot.

Figure 3: Mot segmenté au moyen d'un algorithme (après la suppression de tous les points de segmentation erronés)

2.2 Extraction de caractéristiques

L'extraction de caractéristiques est de fait l'une des deux fonctions essentielles d'un SRC. Elle comprend la mesure des caractéristiques de la forme d'entrée (nettoyée) qui sont pertinentes à la classification. Lorsque l'extraction des caractéristiques est terminée, la forme est représentée par l'ensemble des caractéristiques extraites.



a) boucle, points d'extrémité, un point et une jonction

b) contour d'une lettre « C »

Figure 4 (a et b): Certaines caractéristiques des caractères

Il existe un nombre infini de caractéristiques possibles que l'on peut extraire d'une forme finie à deux dimensions. Il faut toutefois ne s'attarder qu'aux caractéristiques qui ont une pertinence possible pour la classification, ce qui suppose qu'au cours de la période de conception, le spécialiste s'attarde aux caractéristiques qui, selon une certaine technique de classification, apporteront les résultats les plus certains et les plus efficaces.

Ainsi, dans un alphabet à deux symboles, le « 0 » et le « 1 », la hauteur de la donnée d'entrée (chiffre) ne constitue pas une caractéristique de différenciation; elle a donc peu d'importance. Par ailleurs, le nombre d'angles aigus dans la forme constituerait (éventuellement) une caractéristique de différentiation, le « 0 » n'ayant aucun angle aigu et le « 1 » n'en ayant qu'un (dans leur forme imprimée).

Parmi les divers types de caractéristiques proposés dans la documentation, mentionnons :

- Histogrammes horizontaux et verticaux
- Renseignements sur la sphéricité (p. ex. pente) et les extrémités locales de la sphéricité (de la ligne constituant ou correspondant à un mot) [15].
- Caractéristiques topologiques, dont les boucles (un groupe de pixels blancs entourés par des pixels noirs), les points d'extrémité (des

points ayant 1 point environnant seulement), les points (un amalgame de disons 1 à 3 pixels) et les jonctions (points qui comptent plus de 2 points à proximité) - dans des images amincies en noir et blanc.
- Paramètres de fonctions polynomiales (ou autres) d'ajustement des courbes [30].
- Renseignements sur le contour. Si un contour constitue la limite extérieure d'une forme - reportez-vous à la figure 4 (b).

D'un point de vue abstrait, l'extraction des caractéristiques établit la totalité du mappage du modèle de chaque entrée de son système spatial d'origine (p. ex. Euclidien) de coordonnées en un seul point dans un espace de « caractéristiques ». Cet espace est déterminé par les caractéristiques N extraites, et possède donc des dimensions N. Les axes N délimitant l'espace de caractéristiques sont orthogonaux, lorsque les caractéristiques sont indépendantes les unes des autres (dont la hauteur et la largeur d'un chiffre).

En bref, si l'objectif de l'extraction des caractéristiques est d'établir le mappage des modèles d'entrée en des points dans l'espace de caractéristiques (reportez-vous à la figure 5), l'objet de la classification est alors d'attribuer, à chaque point dans l'espace, une étiquette pertinente (p. ex. un « 1 »). C'est donc dire que dès qu'une forme est soumise à un mappage, le problème devient un problème de classification classique.

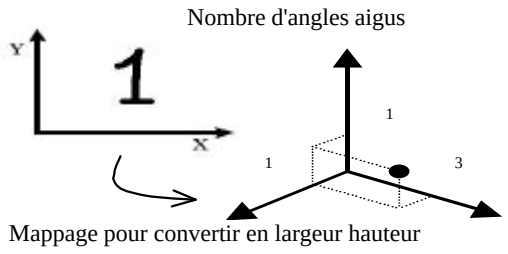


Figure 5 : Extraction de caractéristiques en passant d'un mappage euclidien à un espace de caractéristiques.

2.3 Classification des formes d'entrée

Dès que la forme d'entrée est soumise à un mappage en un point dans l'espace de caractéristiques, on passe ensuite à l'étiquetage de chacun des points. Parmi les techniques de classification, soulignons les :

- Systèmes à base de règles
- Arbres de décision [18]
- Techniques agglomératives [19]
- Réseaux de neurones artificiels
- Modèles de Markov cachés [20]

Toutes ces techniques possèdent une caractéristique commune, elles divisent tout l'espace de caractéristiques en sous-espaces de moindre grandeur, chacun comprenant précisément des points de la même classification. Certaines de ces techniques sont décrites plus en détail ci-dessous.

- **Les systèmes à base de règles** (y compris les systèmes experts) utilisent habituellement les règles SI… ALORS pour établir jusqu'à quel point les conditions dans la partie SI sont conformes au modèle. Dans les systèmes à base de règles, il est possible que deux règles ou plus (comportant des recommandations de classification différentes) soient applicables à la même forme d'entrée. Il en découle des conflits et il faut alors avoir recours à des mécanismes de résolution de conflits. Là encore, dans un système de reconnaissance de caractères comportant un alphabet « 0/1 », on peut s'attendre à avoir une règle telle que : SI (la forme à grande boucle), ALORS (la classe = « 0 »).
- **Les arbres de décision** peuvent être considérés comme une structure d'arbre utilisée pour faciliter la prise de décision. Un arbre possède un seul point d'entrée au sommet, et un nombre indéfini de nœuds feuilles à classe unique à sa base. La méthode ID3 [18] est une méthode très populaire pour ériger automatiquement un arbre de décision à partir d'un ensemble de formes déjà classifiées. Dès qu'un arbre de décision est érigé, il peut servir à classifier de nouvelles formes inconnues.

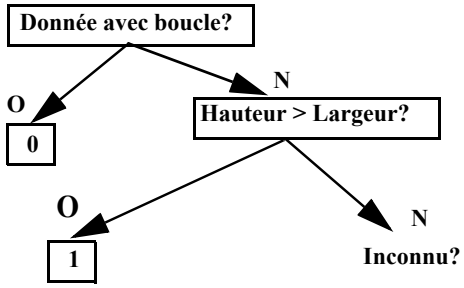


Figure 6 : Arbre de décision simple

- **Les techniques agglomératives** tentent essentiellement de chercher des points dans l'espace de caractéristiques qui sont près les uns des autres et de les placer dans la même classe. Un moyen de le faire consiste à attribuer arbitrairement une classe à chaque point dans un ensemble de points non classés, puis de trouver le centre (ou la moyenne) de chaque groupe de points similaires classés. De là, chaque point est réattribué à la classe du point-centre qui s'en rapproche le plus. S'ensuit une renumérisation des points-centres des différentes classes, et le processus est répété jusqu'à ce qu'il n'y ait plus d'attributions à faire (reportez-vous aux K-moyennes à la note 21)

2.4 Post-traitement

Le post-traitement comprend la vérification, l'exécution de l'action et l'adaptation. L'objectif de la vérification est d'accroître le niveau de confiance dans la classification effectuée; une telle vérification peut être effectuée de diverses façons. L'une de ces façons consiste à utiliser une base de données comportant des combinaisons de 2 ou 3 lettres pour vérifier si la séquence des lettres reconnues ne comprend pas de combinaisons impossibles (p. ex. « zdh »). Une autre possibilité est d'utiliser un dictionnaire pour vérifier si une certaine séquence de caractères constitue un mot valide. En règle générale, cette méthode est moins fiable puisque les mots exacts qui ne sont pas consignés au dictionnaire sont rejetés. En plus des dictionnaires contenant des lettres et/ou des mots, un modèle grammatical formel de niveau plus élevé peut être utilisé pour vérifier l'exactitude d'expressions ou de phrases entières [4].

Outre la vérification, un SRC effectue certaines démarches en réaction à la reconnaissance. Ainsi le système dont il est question à la note [22] tente de déduire certaines caractéristiques personnelles de l'auteur en appliquant une base de règles à l'ensemble de caractéristiques extraites d'un échantillon de son écriture (p. ex. l'élargissement de la marge de gauche indique une tendance à la fatigue à mesure que progresse le travail).

Et plus intéressant encore, certains SRC plus perfectionnés modifient leur propre poids (réseau de neurones artificiels) ou les paramètres probabilistes (modèles de Markov cachés) en vue de s'adapter pour réduire la discontinuité entre la performance visée et la performance réelle, dans le but d'améliorer la performance à venir.

3.0 Conclusions et défis à relever

La problématique de la reconnaissance des caractères constitue un sous-ensemble de la reconnaissance de formes, la reconnaissance des caractères étant limitée aux formes à base de texte. L'objectif de tout système de reconnaissance de caractères est de tirer automatiquement un sens d'une image à deux dimensions (ou d'une trace) d'une entrée de texte. Il existe de nombreux type de SRC. Certains lisent les chèques, d'autres reconnaissent des mots imprimés à partir d'une image balayée. On peut dire, cependant, que tous les SRC comportent quatre parties fonctionnelles : le pré-traitement, l'extraction des caractéristiques, la classification des formes et le post-traitement. Les systèmes de reconnaissance de caractères ne possèdent pas tous ces quatre parties et certains ont des composants supplémentaires. Presque tous doivent cependant arriver à mesurer les caractéristiques et à attribuer, pour chaque forme d'entrée, une classe compréhensible.

Selon les auteurs, quelles que soient les techniques utilisées, toutes les méthodes de reconnaissance de caractères doivent relever deux importants défis : la segmentation et l'adaptation. La segmentation, ou le manque de segmentation, constitue le plus grand problème auquel sont confrontés les concepteurs qui tentent de monter un SRC libre de toute restriction. Le chevauchement des caractères, le chevauchement des mots et la présence des renseignements non utiles (p. ex. le bruit) chevauchant les deux, sont, en fait, des problèmes de segmentation. Il est plutôt difficile de déterminer où débute et où se termine un mot sans au préalable avoir reconnu le caractère ou le mot. La segmentation est cependant souvent nécessaire avant la reconnaissance (ou la classification). Nous sommes donc confrontés au problème de l'œuf ou la poule qui, toujours selon les auteurs, ne peut être réglé qu'en procédant progressivement ou qu'en laissant tomber l'étape de la segmentation.

L'autre problème important dans la reconnaissance de caractères est l'adaptation, particulièrement en l'absence de la rétroaction directe (d'un être humain) pour redresser les erreurs. C'est dire que l'apprentissage ne serait pas surveillé et que l'incertitude persisterait. C'est que la machine aurait à décider elle-même de l'erreur et de l'emplacement de l'erreur, une tâche qui dans les meilleures circonstances peut être ardue. Il y a cependant bon nombre de techniques d'apprentissage automatique offertes dans la documentation proposée [24] qui pourraient être utiles aux chercheurs. La segmentation et l'adaptation sont toutes les deux des problèmes liés de près aux problèmes de niveau 3-1 de Mori [25].

4.0 Références

[1]. S. Di Zenzo et al. (1992) : « Optical Recognition of Hand-printed Characters of any Size, Position, and Orientation », dans IBM Journal of Research & Development. volume 36, no 3.

[2]. Sabri A. Mahmoud (1991) : « Skeletonization of Arabic Characters using Clustering Based Skeletonisation Algorithm (CBSA) », dans Pattern Recognition, volume 24, no 5, pages 453 à 464. Pergamon Press.

[3]. A. C. Downton & S. Impedovo (éditeurs) (1997) : « Progress in Handwriting Recognition », chapitre sur l'extraction des caractéristiques, publié par World Scientific.

[4]. J. J. Hull (1994) : « Language Level Syntactic and Semantic Constraints Applied to Visual Word Recognition », dans « Fundamentals of Handwriting Recognition », par Sebastiano Impedovo (éditeur), publié par Springer-Verlag.

[5]. Sargur Srihari (1997) : « Recent Advances in Off-line Handwriting Recognition at CEDAR », dans A.C. Downton & S. Impedovo (éditeurs) : « Progress in Handwriting Recognition », publié par World Scientific.

[6]. Boris Alexandrovsky (1997) : « A Thalamocortical Algorithm That Performs Handwritten Character Recognition », dans A.C. Downton & S. Impedovo (éditeurs) : « Progress in Handwriting Recognition », publié par World Scientific.

[7]. T. Y. Zhyang & C. Y. Suen (1985) : « A Fast Parallel Algorithm for Thinning Digital Patterns », dans Communications of the ACM, volume 27, no 3.

[8]. M. Ahmed & R. Ward : « A rule-based system for thinning symbols to their central lines », soumis au journal de l'IEEE sur PAMI en juin 1998.

[9]. R. R. Bailey et M. Srinath (1996) : « Orthogonal Moment Features for Use with Parametric and Non-Parametric Classifiers », dans les comptes rendus de l'IEEE traitant sur Pattern Analysis and Machine Intelligence; volume 18, no 4.

[10]. Kauppinen et al (1995) : « An experimental comparison of autoregressive and Fourier-based descriptors in 2D shape classification », dans les comptes rendus de l'IEEE sur Pattern Analysis and Machine Intelligence; volume 17, no 2.

[11]. G. Wilfong et al (1996) : « On-Line Recognition of Handwritten Symbols », dans les comptes rendus de l'IEEE sur Pattern Analysis and Machine Intelligence; volume 18, no 9.

[12]. H. Al-Yousefi et S. S. Udpa (1992) : « Recognition of Arabic Characters » dans les comptes rendus de l'IEEE sur Pattern Analysis and Machine Intelligence; volume 14, no 8.

[13]. U. Mahedavan et S. N. Srihari (1995) : « Gap Metrics for Word Separation in Handwritten Lines», à la 3rd International Conference on Document Analysis and Recognition, Montréal, Canada.

[14]. T. Caesra et al. (1994) : « Handwriting Recognition by Statistical Methods », dans « Fundamentals of Handwriting Recognition », par Sebastiano Impedovo (éditeur), publié par Springer-Verlag.

[15]. X. Li, R. Parizeau, et R. Plamondon (1997) : « Detection of Extreme Points of On-Line Handwritten Scripts », dans A. C. Downton et S. Impedovo (éditeurs): « Progress in Handwriting Recognition », publié par World Scientific.

[16]. H. Beigi (1997) : « Pre-Processing the Dynamics of On-line Handwriting Data, Feature Extraction and Recognition », dans A. C. Downton etS. Impedovo (éditeurs) : « Progress in Handwriting Recognition », publié par World Scientific.

[17]. O. Due Trier et al. (1995) : « Feature Extraction Methods for Character Recognition », dans Pattern Recognition, volume 29, no 4.

[18]. J. R. Quinlan (1986) : « Induction of Decision Trees », dans Machine Learning, 1:81-106.

[19]. http://www.ee.ic.ac.uk/hp/staff/sjrob/ Projects/cluster.html. Dernière consultation, le 15 juin 1999.

[20]. [20] L. R. Rabiner (1989) : « A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition », dans les travaux de l'IEEE, volume 77, no 2.

[21]. K. Fukunaga (1990): «Introduction to Statistical Pattern Recognition», publié par Academic Press.

[22]. G. Sheikholeslami et al (1997) : « Computer Aided Graphology » dans A. C. Downton et S. Impedovo (éditeurs) : « Progress in Handwriting Recognition », publié par World Scientific.

[23]. M. Blumenstein et B. Verma (1998) : « A Neural Based Segmentation and Recognition Technique for Handwritten Words », World Congress on Computational Intelligence, Anchorage, Alaska.

[24]. G. Brisco & T. Caelli (1996) : « A Compendium of Machine Learning, Volume 1: Symbolic Machine Learning », publié par Ablex Publishing Corporation, Norwood, NJ.

[25]. S. Mori (1994) : « Historical Review of Theory and Practice of Handwritten Character Recognition », dans « Fundamentals of Handwriting Recognition », par Sebastiano Impedovo (éditeur), publié par Springer-Verlag.

Au propos des auteurs

Nawwaf Kharma est actuellement chargé de cours au Département de génie électrique et informatique à l'université de Colombie-Britannique à Vancouver où il poursuit des recherches sur la reconnaissance de caractères en ligne et sur l'optimisation des systèmes de reconnaissance de formes. Il a auparavant été professeur adjoint à l'université de Paisley en Écosse où il enseignait l'informatique et le génie logiciel. Sa thèse de doctorat portait sur l'élaboration d'un algorithme d'apprentissage machine incrémental inspiré par la psychologie pour des applications robotiques.

Rabab Ward est professeure au Département de génie électrique et informatique à l'université de Colombie-Britannique et directrice du Centre for Integrated Computer Systems Research (CICSR). Elle est de plus chercheure principale chez Ward Laboratories Inc., une entreprise de la C.-B., dont la mission est de faire passer la technologie du laboratoire à l'industrie. Madame Ward est spécialisée dans le traitement et les applications des signaux numériques à la cablô-diffusion, à la télévision numérique, à la compression vidéo et à l'imagerie médicale, dont la mammographie, la microscopie et les images cellulaires. Détentrice de six brevets, elle a publié plus de 150 articles et écrit des chapitres dans des livres scientifiques.