UNIVERSITE PIERRE ET MARIE CURIE – PARIS VI

INSTITUT DES SYSTEMES INTELLIGENTS ET ROBOTIQUES

HABILITATION A DIRIGER DES RECHERCHES

RECONNAISSANCE DES FORMES : APPRENTISSAGE ET FUSION D'INFORMATIONS

Présentée par

LIONEL PREVOST

Soutenue le 7 décembre 2007 devant le jury composé de :

Isabelle BLOCH	Rapporteur	ENST, Paris
Bernadette DORIZZI	Rapporteur	INT, Evry
Simone MARINAI	Rapporteur	Université de Florence
Philippe BIDAUD	Examinateur	Université Paris VI
Laurent HEUTTE	Examinateur	Université de Rouen
Maurice MILGRAM	Examinateur	Université Paris VI

Résumé

Ce mémoire présente une synthèse des travaux que j'ai menés dans le domaine de la reconnaissance des formes. D'un point de vue méthodologique, je me suis intéressé à toutes les phases du processus : analyse de données pour la sélection de caractéristiques, apprentissage automatique pour la classification ou la détection et fusion d'informations. Concernant cette dernière, j'ai privilégié les approches basées sur une combinaison de classifieurs multiples. J'ai vérifié expérimentalement l'efficacité du paradigme « diviser pour mieux régner », que j'ai décliné de plusieurs manières : en entraînant des classifieurs sur des représentations alternatives de données et en les combinant, mais aussi en partitionnant l'espace de représentation et en y déployant des classifieurs spécialisés. L'effort a porté tant sur les architectures que sur les opérateurs de fusion, en privilégiant toujours les solutions les plus efficaces et les plus génériques. Mes derniers travaux portent sur les techniques de dopages (*boosting*) qui, en combinant élégamment les phases d'extraction de caractéristiques, de classification et de fusion d'informations, offrent donc un cadre idéal pour la résolution de problèmes complexes en reconnaissance des formes.

Plusieurs applications sont présentées pour valider ces approches. Les recherches doctorales effectuées en reconnaissance de caractères manuscrits isolés se sont poursuivies et ont naturellement conduit à s'intéresser à la reconnaissance de textes manuscrits dans les dispositifs nomades. Par la suite, des algorithmes de localisation et de poursuite du visage et des caractéristiques faciales ont vu le jour, pour la biométrie et l'interaction homme-machine. A la faveur d'un contrat industriel, nous développons actuellement des méthodes de détection de véhicules dans les images de scènes routières avec détermination du type de véhicule détecté. Enfin, dans le cadre d'un projet d'assistance à la personne, nous avons entrepris des recherches sur la détection et la reconnaissance des textes dans les images, en vue d'offrir aux non-voyants une perception plus riche de leur environnement.

Abstract

This thesis presents a synthesis of my research works in the field of pattern recognition. From a methodological point of view, I was interested in all the phases of the process: data analysis for feature selection, machine learning for classification or detection and information fusion. Considering the latter, I privileged approaches based on a combination of multiple classifiers. I verified experimentally the efficiency of the paradigm "divide and conquer". I used it in different ways: by training classifiers on alternative data representations and then combining them, and also partitioning the feature space and deploying specialized classifiers. Efforts are made to find fusion architectures and fusion operators always keeping in view effective and generic solutions. My recent works are related to boosting techniques which combine elegantly the stages of feature extraction, classification and information fusion and offer an ideal framework for the solution of complex problems in pattern recognition.

Several applications are presented to validate these approaches. My doctoral researches, carried out in the domain of handwritten character recognition, naturally drove to a handwritten text recognition module for handheld devices. Thereafter, algorithms for localization and tracking of face and facial features were developed, for biometrics and man-machine interaction. With the help of an industrial contract, we are currently developing methods for vehicles detection in road scenes including recognition of their type. We are currently putting effort to a camera-based text detection and recognition system, in order to help the visually impaired persons to understand their environment in a better way.

TABLE DES MATIERES

TABLE DES FIGURES

1. NOTICE INDIVIDUELLE

1.1 ETAT CIVIL

LIONEL PREVOST
Né le 11 juin 1970
Célibataire
Nationalité française

29 rue des laitières
94300 Vincennes
(33)1.43.74.44.35
prevost.lionel@free.fr

Maître de conférences, 61ème section
Université Pierre et Marie Curie (UPMC)
Institut des Systèmes Intelligents et Robotiques

3 rue Galilée
94200 Ivry sur Seine
(33)1.44.27.23.48
lionel.prevost@upmc.fr

1.2 FORMATION

1988 Baccalauréat D

1991 DEUG A (mention assez Bien) UPMC

1993 Maîtrise de Physique Appliquée (mention assez Bien) UPMC

1994 DEA d'électronique, option électronique et instrumentation (mention assez Bien) UPMC, stage réalisé à l'ESPCI sous la direction de G. Dreyfus

1998 Doctorat d'Informatique (mention très honorable avec les **félicitations du jury**) UPMC

« Reconnaissance de l'écrit dynamique : applications à l'analyse des formules mathématiques manuscrites »

Directeur de Thèse : M. Milgram (Pr)

Jury : H. Emptoz (Président), B. Dorizzi & G. Lorette (Rapporteurs), C. Faure, P. Gallinari, P. Garda & M. Milgram.

1.3 DATES IMPORTANTES

1998 Attaché Temporaire d'Enseignement et de Recherche (ATER)
UPMC (UFR 924 : EEA et Applications de la Physique).

2000 Maître de Conférences (61ème section)
UPMC (UFR 924 : EEA et Applications de la Physique).

2003 Bénéficiaire de la **Prime d'Encadrement Doctorale et de Recherche**

1.4 ACTIVITES DE RECHERCHE

Thématiques :

D'un point de vue méthodologique, mes travaux couvrent toutes les étapes du processus de reconnaissance des formes :

- codage : partant d'un signal monodimensionnel, bidimensionnel (image) ou multidimensionnel (vecteur de descripteurs), je cherche à sélectionner les caractéristiques les plus pertinentes de ce signal par des méthodes d'analyse de données (analyse en composante principale, analyse discriminante …) et/ou d'apprentissage automatique.

- classification : après la phase d'extraction précédente, j'ai développé plusieurs algorithmes de classification utilisant des méthodes génératives (statistiques ou neuronales), des méthodes discriminantes et des méthodes cumulatives, afin de traiter des problèmes complexes : forte variabilité, nombre de classes élevé … Les problèmes de détection (caractérisés par l'absence d'exemples « négatifs ») ont aussi été largement abordés.

- fusion d'informations : les limites des méthodes précédentes et l'augmentation de puissance de calcul des ordinateurs m'ont conduit à développer des stratégies de combinaison de classifieurs traitant des représentations alternatives des données initiales. L'influence du comportement des classifieurs de bases (indépendance, conflit), des architectures mises en œuvre (parallèle, hiérarchique, hybride) et des opérateurs de fusion (statistique, neuronal, flou) a été analysée en détail.

- dopage : Je travaille actuellement sur les méthodes de *boosting* qui visent à intégrer en un unique processus les phases de codage, classification et fusion d'informations.

Applications :

Les méthodes précédentes sont évaluées sur différents problèmes réels. Chronologiquement, j'ai effectué des recherches sur :

- la reconnaissance d'écriture (1994 – 2004) : dans le cadre de mes recherches doctorales et de la thèse de Loïc Oudot, j'ai développé des méthodes de reconnaissance de caractères puis de lecture de textes « dynamiques » c'est-à-dire écrits à l'aide d'un stylo électronique sur un écran tactile. Ces systèmes étaient utilisés sur des dispositifs « nomades » (ordinateurs, assistants personnels, *smartphones*).

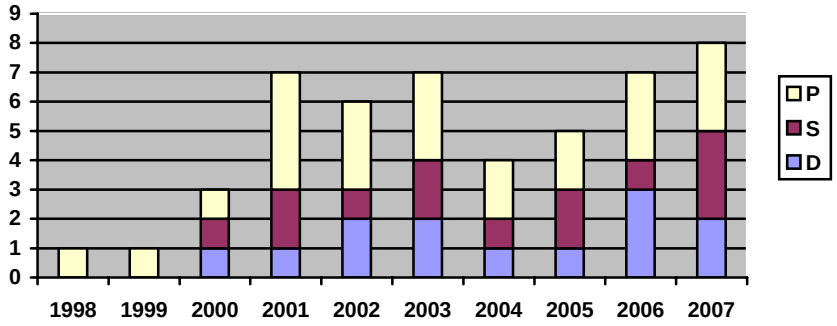
- l'analyse du visage humain (depuis 2002) : nous avons initié, avec la thèse de Rachid Belaroussi, des recherches sur la localisation et le suivi de visages, dans les images et les vidéos, pour la biométrie et l'interaction homme-machine « intelligente ». Je participe aussi à des recherches sur la localisation des caractéristiques faciales pour la détermination de l'orientation du visage et sur la synthèse réaliste.

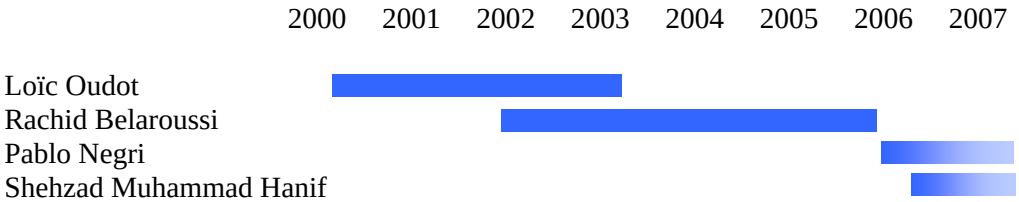
- La détection de véhicules (depuis 2006) : à la faveur d'un contrat avec PSA, la thèse de Pablo Negri porte sur l'analyse d'algorithmes de détection de véhicules dans des scènes routières et la détermination du type de véhicule détecté. L'objectif est de développer des fonctions d'assistance à la conduite (*adaptive cruise control, stop and go, collision warning* …) et de réduire, à terme, le nombre d'accidents de voiture.

Co-encadrements :

- **4** thèses (dont 2 soutenues).
- **13** stages de DEA/DESS/Master 2ème année (4 mois à temps plein).
- **24** projets de Maîtrise/Master 1ère année, DEA/DESS/Master 2nde année, Ingénieur fin d'étude (1 Mois à temps plein ou équivalent).

Encadrement : bilan 1998-2007.

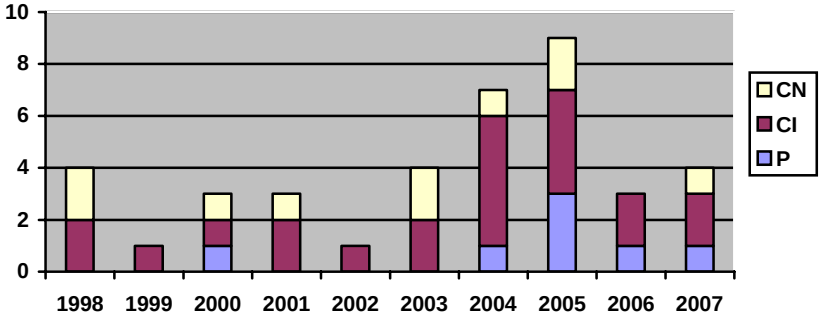




Publications :

- **1** Direction d'Ouvrage.
- **7** Publications dans des revues avec comité de lecture.
- **25** Communications dans des Conférences Internationales avec actes et comité de lecture.
- **12** Communications dans des Conférences Nationales avec actes et comité de lecture.

Publications : bilan 1998 – 2007.



1.5 ANIMATION SCIENTIFIQUE

- **Organisateur et Président du comité de programme** du workshop ANNPR (Artificial Neural Networks in Pattern Recognition) de l'IAPR-TC3 (International Association for Pattern Recognition – Neural Networks & Computational Intelligence) qui se tiendra à Paris du 2 au 4 juillet 2008 sous le haut patronage de l'IAPR.

- **Membre du comité de programme**
  Colloque International Francophone sur l'Ecrit et le Document : CIFED 2004 et CIFED 2006
  International Workshop/Conference on Frontier of Handwriting Recognition: IWFHR 2006 et ICFHR 2008
  International Conference on Artificial and Computational Intelligence for Decision, Control and Automation: ACIDCA-ICMI 2005

- **Membre du comité d'organisation** des conférences CIFED 2006, HANDICAP 2006.

- **Président de session**
  Conférence ANNPR 2006, session 1 « Visual Object Recognition »
  Conférence CIFED 2006, session 7 « Ecriture et documents en ligne »

- **Relecteur pour les revues** Pattern Recognition (1 article), Pattern Recognition Letters (1 article), IEEE Systems, Man & Cybernetics (1 article), Traitement du Signal (2 articles).

- **Relecteur pour les conférences** CIFED 2004 (4 communications), ACIDCA-ICMI 2005 (3 communications), CIFED 2006 (4 communications), IWFHR 2006 (10 communications), ICANN 2006 (1 communication), KES 2007 (1 communication), ICFHR 2008.

- **Membre** de l'IAPR-TC3 (International Association for Pattern Recognition – Neural Networks & Computational Intelligence), de l'IAPR-TC11 (International Association for Pattern Recognition – Reading Systems), de l'ISIF (International Society of Information Fusion), du GRCE (Groupe de Recherche en Communication Ecrite), des GdR-ISIS et GdR-I3.

- **Participation à des jurys de thèse** :
  Encadrant : Loic Oudot (UPMC, décembre 2003)
  Encadrant : Rachid Belaroussi (UPMC, décembre 2006)
  Président du jury : Emine Krichen (INT, octobre 2007)
  Examinateur : William Ivaldi (UPMC, fin 2007).

- **Membre du comité de suivi** de la thèse de Ludovic Simon (LCPC).

1.6 PARTICIPATION A DES CONTRATS INDUSTRIELS

1. SAGEM

Objectif : Restituer une vue de face d'une personne en mouvement à partir d'une série de vues multi-caméras
Financement : SAGEM Eragny
Montant : 45 000€
Dates : Mars 2004 - Février 2007
Doctorant : W. Ivaldi
Statut : participant

2. PSA

Objectif : Détecter et classer par type les véhicules à l'avant du véhicule porteur du système d'acquisition et de traitement. Conditions diurnes et routes planes, éclairage et visibilité variables.
Financement : PSA
Montant : 78 000€
Dates : Mars 2006 - Février 2008
Doctorant : P. A. Negri
Statut : co-responsable avec X. Clady

1.7 ACTIVITES D'ENSEIGNEMENT

J'assure actuellement 192 heures (équivalent TD) d'enseignement réparties comme suit :

CM/TD Automatique (EPU ELI 4, Master SDI1).

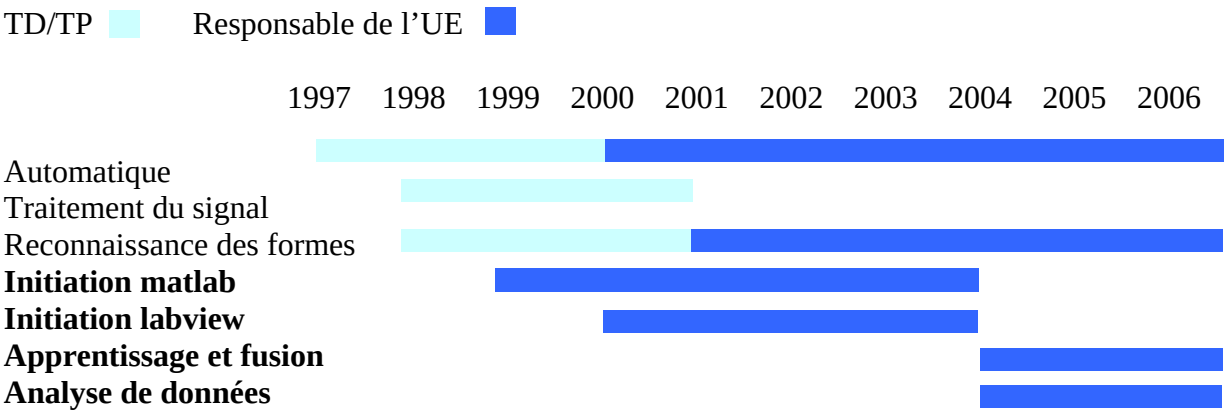
CM/TP Reconnaissance des Formes (Master SDI2)

CM/TP Méthodes Connexionnistes, Apprentissage et Fusion (Master SDI2)

CM/TP Analyse et Codage des signaux (Master SDI2)

Projets Master SDI1 & 2

Enseignement : bilan 1997-2006.



(Les cours que j'ai montés sont indiqués en gras)

1.8 RESPONSABILITES ADMINISTRATIVES

Depuis 2006 :
Directeur des études (avec H. Kokabi, Pr) à Polytech'Paris-UPMC : Parcours des Ecoles d'Ingénieurs Polytech' (PeiP : classe préparatoire intégrée de l'école).

Missions :

- recrutement : examen des dossiers, organisation des entretiens (plus de 400 candidats)
- organisation : emplois du temps, coordination des équipes pédagogiques, présidence de jury
- suivi : 120 étudiants en première année à la rentrée 2007
- liaison : avec le réseau Polytech et le cycle ingénieur (années 3 à 5)

Depuis 2004 :
UE Analyse et Codages des Signaux (Master SDI2)
UE Initiation à la Reconnaissance des Formes (Master SDI2)
UE Méthodes connexionnistes, Apprentissage et Fusion d'Informations (Master SDI2)

2002-2004 :
UE Formation Générale (Licence professionnelle LIOVIS)
UE Visualisation2 (Licence professionnelle LIOVIS)

1.9 FONCTIONS ELECTIVES

Membre de droit du Conseil de Direction de Polytech-Paris-UPMC depuis 2006.

Membre élu de la CS 61ème section de L'Université Pierre et Marie Curie depuis 2004.

Membre nommé de la CS 61ème section de L'Université de Rouen depuis 2004.

Membre élu du Conseil de Laboratoire du LISIF de 1996 à 2004.

1.10 ENCADREMENTS

Doctorats :

1. Oudot L., Fusion d'informations et adaptation pour la reconnaissance de textes manuscrits dynamiques
Bourse MESRT
Directeur de thèse co-encadrant : M. Milgram (25%)
Thèse débutée en septembre 2000 et soutenue le 4 décembre 2003
Jury : Vincent N. (Présidente), Lorette G. & Paquet T. (Rapporteurs), Gallinari P., Knerr S., Milgram M. & Prevost L. (Examinateurs)
Mention Très Honorable
Actuellement Ingénieur R&D SAGEM division Biométrie.

2. Belaroussi R., Contribution à la détection et la localisation de visages dans les images
Bourse MESRT
Directeur de thèse co-encadrant : M. Milgram (75%)
Thèse débutée en décembre 2002 et soutenue le 21 décembre 2006
Jury : Devars J. (Président), Bollon P. & Postaire J.G. (Rapporteurs), Davoine F., Garcia C., Milgram M., Prevost L. & Séguier R. (Examinateurs)
Mention Très Honorable
Actuellement Post-Doctorant à l'UPMC.

3. Negri P. A., Méthodes algorithmiques pour la détection et la classification de véhicule
Bourse CIFRE (contrat avec PSA)
Directeur de thèse : M. Milgram ; co-encadrant : X. Clady (60%).
Thèse débutée en mars 2006, soutenance prévue : mars 2008

4. Muhamad Hanif S., Aide aux déficients visuels, localisation et interprétation de l'information textuelle et symbolique contenue dans l'environnement
Bourse SFERE[1],
Directeur de thèse : M. Milgram
Thèse débutée en septembre 2006, soutenance prévue : septembre 2009

Stages (DEA/DESS/Master 2 : 4 mois à temps plein)

1. Oudot L., Reconnaissance de textes manuscrits dynamiques (DEA 2000).

2. Dewaele G., Reconnaissance de formules mathématique manuscrites : Développement d'un logiciel pédagogique d'aide à l'apprentissage du calcul (DEA 2001).

3. Pajadon A., Reconnaissance de textes manuscrits dynamiques : Méthodes statistique et lexicale d'adaptation au scripteur (DEA 2001).

4. Torkamanlou P., Méthodes pour la localisation de visages dans les images (DEA 2002).

5. Das Neves J. C., Localisation et reconnaissance de plaques minéralogiques (DEA 2003).

6. Moises A., Apprentissage adaptatif pour la reconnaissance de textes manuscrits dynamiques, (DEA 2003).

[1] Société Française d'Exportation des Ressources Educatives

7. Michel-Sendis C., Exploration de la voie phonologique en reconnaissance de textes manuscrits (DESS 2004).

8. Chetouani A., Authentification de personnes par analyse du visage (Master2, 2005).

9. Eshkevari B., Détermination de l'orientation d'un visage (Master2, 2005).

10. Muhamad S. H., Réseaux de neurones à convolution pour la localisation de caractéristiques faciales (Master2, 2006).

11. Leroy O., Localisation de points caractéristiques dans un visage (Master2, 2007).

12. Nguepdjop L., Conception d'un système d'aide au déplacement des déficients visuels : détection de textes dans les images (Master2, 2007).

13. Belaidi A. A., Détection de véhicules (Master2, 2007).

Projets (Ingénieur/Maîtrise/Master 1/DEA/DESS/Master 2 : 1 mois à temps plein ou équivalent)

1. Da Silva I., Extraction d'objets dans une formule mathématique manuscrite (MEEA 1998).

2. Moungala S. & Smaili M., Authentification de scripteurs : caractérisation de la signature (MEEA 1999).

3. Jangal F. & Minot M., Fusion de données par réseaux de neurones en présence de données manquantes ou incertaines (MEEA 2000).

4. Courtois F. & Meunier C., Analyse d'opérations manuscrites pour l'aide à l'apprentissage du calcul (DEA 2001).

5. Henri P. & Pajadon A., Paiement sécurisé : Authentification de scripteurs (DEA 2001).

6. Tazerout R. & Gouge G., Localisation automatique des zones de texte dans une image (DESS 2001).

7. Alvarez C. & Fernandez L., Identification et commande d'un système de régulation de température par réseau de neurones (MEEA 2001).

8. Martos L. & Torkamanlou P., Cours HTML : les réseaux de neurones artificiels (MEEA 2001).

9. Robert S. & Tricquet M., Reconnaissance de l'écriture : auto-classification (MEEA 2002).

10. Michel-Sendis C. & Moises A., Coopération modélisation-discrimination pour la reconnaissance de l'écrit (MEEA 2002).

11. Bloquet N., Gelblat M. & Viard R., Reconnaissance de l'écrit dynamique : adaptation au scripteur (DESS 2002).

12. Barthet M. & Boumriga A., Classifieur neuronal évolutif : application à la reconnaissance des phonèmes (IST 2003).

13. Bertin A. & Delpech A., Paiement sécurisé : authentification de scripteur sur code chiffré (MEEA 2003).

14. Chang R. & Para E., Reconnaissance de visages par réseaux de neurones (MEEA 2003).

15. Chetouani A. & Qu X. (MEEA 2004).

16. Benfares D. & Hamoudi H., Réseaux de neurones hybrides pour la localisation des traits caractéristiques du visage (MEEA 2004).

17. Adamo H. & Traoré A., Apprentissage statistique évolutif (Master2 2005).

18. Li C., Wang Y. & Yao Y., Modélisation statistique de la teinte chair (Master1 2005).

19. Bourre-Sumimoto R. & Falina E., Réseaux auto-associatifs hybrides pour la localisation des caractéristiques faciales (Master2 2006).

20 Medjahed H., Mouhou T. & Muhamad S., Réseaux à convolution pour la localisation des caractéristiques faciales (Master2 2006).

21. Hasting P. & Roussel J., Réseaux auto-associatifs hybrides pour la localisation des points caractéristiques d'un véhicule (Master2 2006).

22. Chautems N. & Leroy O., Analyse du contexte dans une image de véhicule (Master2 2007).

23. Lefranc P.A., Localisation de points caractéristiques dans une image de véhicule (Master2 2007).

24. Ait-Mohand K., Localisation fine des lèvres dans une image de visage (Master2 2007).

1.11 PUBLICATIONS

Direction d'ouvrage :

[1] **Prevost L.**, Marinai S. & Schwenker F., Proceedings of the 3rd ANNPR (IEEE/IAPR-TC3 International Workshop on Artificial Neural Networks in Pattern Recognition), **Lecture Notes in Artificial Intelligence**, à paraître, 2008.

Publications dans des revues internationales avec comité de lecture :

[2] **Prevost L.** & Milgram M., Modelizing character allographs in omni-scriptor frame: a new non-supervised algorithm, **Pattern Recognition Letters**, 21(4), pp 295-302, 2000.

[3] **Prevost L.** & Oudot L., Self-supervised adaptation for on-line script text recognition, **Electronic Letters on Computer Vision and Image Analysis**, Special issue on Document Analysis, 5(2), pp 87-97, 2005.

[4] **Prevost L.**, Oudot L., Moises A., Michel-Sendis C. & Milgram M., Hybrid generative/discriminative classifier for unconstrained character recognition, **Pattern Recognition Letters**, Special issue on Artificial Neural Networks in Pattern Recognition, 26(12), pp 1840-1848, 2005.

[5] Belaroussi R., **Prevost L.** & Milgram M., Algorithm fusion for face localization, **Journal of Advances in Information Fusion**, 1(1), pp 27-38, 2006.

[6] Muhammad Hanif S., **Prevost L.**, Belaroussi R. & Milgram M., Real-time facial feature localization by combining space displacement neural networks, **Pattern Recognition Letters**, Special issue on Pattern Recognition in Multidisciplinary Perception and Intelligence, accepté.

Publications dans des revues francophones avec comité de lecture :

[7] Oudot L. & **Prevost L.**, Techniques de coopération pour la reconnaissance d'écriture en contexte, **Revue d'Intelligence Artificielle**, 18(3), pp 367-382, 2004.

[8] Oudot L., **Prevost L.** & Milgram M., Fusion d'informations et adaptation pour la reconnaissance de textes manuscrits dynamiques, **Traitement du Signal**, 22(3), pp 239-248, 2005.

Communications dans des conférences internationales avec actes et comité de lecture :

[9] **Prevost L.** & Milgram M., Static and dynamic classifier fusion for character recognition, **ICDAR'97** (International Conference on Document Analysis and Recognition), (2), pp 499-506, Ulm, Allemagne, 1997.

[10] **Prevost L.** & Milgram M., Automatic Allograph Selection and Multiple Expert Classification for Totally Unconstrained Handwritten Character Recognition, **ICPR'98** (International Conference on Pattern Recognition), (1), pp 381-383, Brisbane, Australie, 1998.

[11] **Prevost L.** & Milgram M., Non-supervised Determination of Allograph Sub-classes for On-line Omni-scriptor Handwriting Recognition, **ICDAR'99** (International Conference on Document Analysis and Recognition), pp 438-441, Bangalore, Inde, 1999.

[12] **Prevost L.**, Gentric S. & Milgram M., Cooperation and modularity: Application to handwriting character recognition, **ICIF'00** (International Conference on Information Fusion), (2 ThC5); pp 3-8 ; Paris, 2000.

[13] Gentric S., **Prevost L.** & Milgram M., Dynamic handwriting recognition based on an evolutionary neural classifier, **ICANNGA'01** (International Conference on Artificial Neural Networks & Genetic Algorithms), pp 394-396, Pragues, République Tchèque, 2001.

[14] Oudot L., **Prevost L.** & Milgram M., Dynamic Recognition in the omni-writer frame: Application to the hand-printed text recognition, **ICDAR'01** (International Conference on Document Analysis and Recognition), pp 1035-1039, Seattle, Etats Unis, 2001.

[15] **Prevost L.**, Michel-Sendis C., Moises A., Oudot L. & Milgram M., Combining model-based and discriminant classifiers : application to handwritten character recognition, **ICDAR'03** (International Conference on Document Analysis and Recognition), pp 31-35, Edimbourg, Ecosse, 2003.

[16] **Prevost L.**, Moises A., Michel-Sendis C., Oudot L. & Milgram M., A new hybrid adaptive classifier for unconstrained character recognition, **ANNPR'03** (IEEE/IAPR-TC3 International Workshop on Artificial Neural Networks in Pattern Recognition), pp 26-32, Florence, Italie, 2003.

[17] Oudot L., **Prevost L.** & Milgram M., An activation-verification model for on-line text recognition, **IWFHR'04** (International Workshop on Frontier of Handwriting Recognition), pp 9-13, Tokyo, Japon, 2004.

[18] Oudot L., **Prevost L.** & Moises A., Self-supervised adaptation for handwritten text recognition, **IWFHR'04** (International Workshop on Frontier of Handwriting Recognition), pp 485-491, Tokyo, Japon, 2004.

[19] Oudot L., **Prevost L.** Moises A. & Milgram M., Self-supervised writer adaptation using perceptive concepts: application to on-line text recognition, **ICPR'04** (International Conference on Pattern Recognition), (2), pp 598-601, Cambridge, Royaume Uni, 2004.

[20] Belaroussi R., **Prevost L.** & Milgram M., Classifier combination for face localization in color images, **ICIAP'05** (International Conference on Image Analysis and Processing), Lecture Notes in Computer Sciences, pp 1043-1050, Cagliari, Italie, 2005.

[21] Belaroussi R., **Prevost L.** & Milgram M., Model-based classifiers fusion for face and eyes localization, **ICIF'05** (International Conference on Information Fusion), CD D7-3, Philadelphie, Etats-Unis, 2005.

[22] Belaroussi R., **Prevost L.** & Milgram M., Combination of multiple detectors for face and eyes localization, **ISPA'05** (International Symposium on Image and Signal Processing and Analysis), pp 24-30, Zagreb, Croatie, 2005.

[23] Belaroussi R., **Prevost L.** & Milgram M., Combining model-based classifiers for face localization, **MVA'05** (Machine Vision & Applications), pp 290-293, Tokyo, Japon, 2005.

[24] Milgram M., Belaroussi R. & **Prevost L.**, Multi-stage combination of geometric and colorimetric detectors for eyes localization, **ICIAP'05** (International Conference on Image Analysis and Processing), Lecture Notes in Computer Sciences, pp 1010-1017, Cagliari, Italie, 2005.

[25] **Prevost L.**, Belaroussi R. & Milgram M., Multiple neural networks for facial feature localization in orientation-free images, **ANNPR'06** (IEEE/IAPR-TC3 International Workshop on Artificial Neural Networks in Pattern Recognition), Lecture Notes in Artificial Intelligence, (4087), pp 188-197,Reisensburg, Allemagne, 2006.

[26] Muhammad Hanif S., **Prevost L.**, Belaroussi R. & Milgram M., A Neural Approach for Real Time Facial Feature Localization, **SSD'07** (IEEE International Multi-Conference on Systems, Signals and Devices), Hammamet, Tunisie, 2007.

[27] Muhammad Hanif S., & **Prevost L.**, Texture-based text detection in natural scene images: a help to blind and visually impaired persons, **CVHI'07** (Conference and workshop on assistive technologies for people with Vision and Hearing Impairments), à paraître, Grenade, Espagne, 2007.

[28] Muhammad Hanif S., & **Prevost L.**, Text detection in natural scenes images using spatial histograms, **CBDAR'07** (Workshop on camera-Based Document Analysis and Recognition, ICDAR' satellite), à paraître, Curitiba, Brésil, 2007.

[29] Negri P. A., Clady X. & **Prevost L.**, Benchmarking Haar and Histograms of Oriented Gradients features applied to vehicle detection, **ICINCO'07** (IFAC International Conference on Informatics in Control, Automation and Robotics), pp 359-364, Angers, France, 2007.

Communications dans des conférences internationales francophones avec actes et comité de lecture :

[30] **Prevost L.** & Milgram M., Reconnaissance automatique de l'écriture scripte en mode omni-scripteur : un premier pas dans la conception d'un analyseur d'équations, **CIFED'98** (Colloque International Francophone sur l'Ecrit et le Document), pp 364-373, Québec, Canada, 1998.

[31] Oudot L., **Prevost L.** & Milgram M., Système de lecture multi-contextuel pour la reconnaissance de textes manuscrits dynamiques, **CIFED'02** (Colloque International Francophone sur l'Ecrit et le Document), pp 335-344, Hammamet, Tunisie, 2002.

[32] Oudot L., **Prevost L.** & Milgram M., Un modèle d'activation-vérification pour la lecture de textes manuscrits dynamiques, **CIFED'04** (Colloque International Francophone sur l'Ecrit et le Document), pp 117-122, La Rochelle, France, 2004.

[33] Oudot L., **Prevost L.** & Moises A., Techniques d'adaptation au scripteur pour la lecture de textes manuscrits dynamiques, **CIFED'04** (Colloque International Francophone sur l'Ecrit et le Document), pp 211-216, La Rochelle, France, 2004.

Communications dans des conférences nationales avec actes et comités de lecture :

[34] Schwenk H., **Prevost L.** & Milgram M., Comparaison de la distance élastique et de la distance tangente en reconnaissance de caractères on-line, **RFIA'96** (Reconnaissance des Formes et Intelligence Artificielle), (2), pp 589-596, Rennes, 1996.

[35] **Prevost L.** & Milgram M.; Coopérations pour la reconnaissance de caractères dynamiques isolés, **RFIA'98** (Reconnaissance des Formes et Intelligence Artificielle), (3), pp 233-240, Clermont Ferrand, 1998.

[36] **Prevost L.** & Milgram M.; Détermination automatique des allographes pour la reconnaissance de l'écriture manuscrite dynamique en mode omni-scripteur, **JFA'98** (Journées Françaises sur l'Apprentissage automatique), pp 150-161, Arras, 1998.

[37] Gentric S., **Prevost L**. & Milgram M., Un classifieur neuronal évolutif dédié à la génération automatique de sous-classes. Application à la reconnaissance d'écriture dynamique, **CAP'00** (Colloque francophone sur l'APprentissage automatique), pp 147-156, Saint-Etienne, 2000.

[38] Oudot L., **Prevost L.** & Milgram M., Reconnaissance de l'écrit dynamique, application à l'analyse de textes, **GRETSI'01** (Traitement du Signal et Images), CD 154, Toulouse, 2001.

[39] Milgram M., Belaroussi R. & **Prevost L.**., Détection de visages sur des images fixes par combinaison de classifieurs discriminants et de modèles, **GRETSI'03** (Traitement du Signal et Images), **pp/CD ???**, Paris, 2003.

[40] **Prevost L.**, Moises A., Michel-Sendis C., Oudot L. & Milgram M., Coopération modélisation-discrimination pour la reconnaissance d'écriture manuscrite, **GRETSI'03** (Traitement du Signal et Images), **pp/CD ???**, Paris, 2003.

[41] Oudot L., **Prevost L.**, & Milgram M., Fusion probabiliste d'informations appliquée à la reconnaissance de l'écrit dynamique, **RFIA'04** (Reconnaissance des Formes et Intelligence Artificielle), (3), pp 1303-1312, Toulouse, 2004.

[42] **Prevost L.**, Oudot L., Moises A., Michel-Sendis C., & Milgram M., Un classifieur hybride incrémental pour la reconnaissance de caractères non contraints, **RFIA'04** (Reconnaissance des Formes et Intelligence Artificielle), (2), pp 789-796, Toulouse, 2004.

[43] Belaroussi R., **Prevost L.** & Milgram M., Combinaison de classifieurs pour la localisation de visage, **GRETSI'05** (Traitement du Signal et Images), pp 941-944, Louvain, Belgique, 2005.

[44] Milgram M. **Prevost L.** & Belaroussi R., Une nouvelle transformation pour la localisation des yeux dans une image de visage monochrome, **GRETSI'05** (Traitement du Signal et Images), pp 671-674, Louvain, Belgique, 2005.

[45] Muhammad Hanif. S., **Prevost L.**, Belaroussi R. & Milgram M., Combinaison de réseaux neuronaux pour la localisation des caractéristiques faciales dans des images de visages sans contrainte d'orientation, **CAP'07** (Conférence francophone sur l'APprentissage automatique), pp 305-307, Grenoble, 2007.

Séminaires et démonstrations

Oudot L., **Prevost L.** & Moises A., Reconnaissance de textes dynamiques : adaptation au scripteur, Journée Jeunes Chercheurs GRCE, Tours, 2003.

Belaroussi R., **Prevost L.** Ivaldi W. & Milgram M., Localisation du visage, des yeux et fusion multi-caméras, **GDR-ISIS**, Journée "Visages", ENST, Paris, 2005.

Belaroussi R., **Prevost L.**, Localisation du visage et des yeux, **Supelec, Rennes**, 2005.

Belaroussi R., **Prevost L.** & Milgram M., A webcam application: face tracking and facial feature detection, **ECCV'06** (European Conference on Computer Vision), Graz, Autriche, 2006.

Muhammad Hanif S., **Prevost L.**, Real-time facial feature localization by combining space displacement neural networks, Journée Jeunes chercheurs "Visage-Geste-Mouvement" ENST, Paris, 2006.

Rapports de contrats industriels

Ivaldi W., Etat de l'art, 2004.

Negri P., Détection de véhicules : état de l'art, 2006.

2. INTRODUCTION

2.1 PARCOURS DE RECHERCHE

Durant toutes ces années, les **réseaux de neurones** ont constitué l'un des fils conducteurs de mes travaux. J'ai fait mes premiers pas dans le monde de la recherche pendant mon DEA en 1994. Mon stage, réalisé au Laboratoire d'Electronique de l'Ecole Supérieure de Physique chimie Industrielle, avait pour objet d'étudier les propriétés des réseaux de neurones artificiels comme outils d'identification et de commande de processus.

Fin 1994, j'ai commencé mon doctorat au Laboratoire Perception, Automatique et Réseaux Connexionnistes de l'Université Pierre et Marie Curie – Paris VI. J'ai mis à profit mon « expertise » sur un problème nouveau pour moi : la reconnaissance des formes, plus exactement la **reconnaissance de caractères manuscrits**. Constatant les limites des méthodes neuronales, je me suis tourné vers d'autres algorithmes de **classification automatique**. Les outils que j'ai développés alors s'inspiraient des algorithmes d'apprentissage supervisé et non supervisé. Leurs performances étaient plus qu'honorables, mais ne suffisaient pas dans la compétition engagée à l'échelle internationale dans ce domaine. Aussi ai-je développé d'autres méthodes de classification alternatives, et surtout des algorithmes de **fusion d'informations** de classification afin de créer une synergie entre ces méthodes ; synergie conduisant à l'émergence d'une solution bien plus robuste. J'ai alors vérifié que les réseaux neuronaux étaient – tout comme les méthodes statistiques – de puissants outils d'intégration d'informations. Par la suite, et toujours dans le même contexte applicatif, j'ai montré que, si leurs performances sur un problème très complexe étaient limitées, ils pouvaient être aussi utilisés comme outils de décision locaux suivant le principe « diviser pour mieux régner ». La solution proposée alors, une combinaison de classifieurs génératif et discriminant, n'était plus du tout guidée par les données (*data driven*), mais bien générique et parfaitement transposable à n'importe quel problème de classification.

Les travaux précédents ont naturellement conduit à s'intéresser à la **lecture automatique de textes** manuscrits. C'était le sujet de la thèse de Loïc Oudot, commencée en 2000 lors de mon recrutement comme Maître de Conférences (section 61) au Laboratoire des Instruments et Systèmes d'Ile de France de l'Université Pierre et Marie Curie. Suite logique donc, mais aussi particulièrement complexe. Il fallait en effet réussir à décomposer le texte manuscrit en unités élémentaires et caractériser celles-ci avant de pouvoir les traiter à l'aide des classifieurs développés pendant ma thèse. L'analyse des modèles de lecture automatique proposés en psychologie perceptive a conduit à l'utilisation de réseaux de neurones comme experts d'analyse et de caractérisation du « signal » (en fait, le tracé du stylo). C'est toutefois le cadre statistique qui a été choisi pour l'intégration de toutes ces informations. Les stratégies de fusion développées précédemment s'appliquaient à des données homogènes (typiquement, des sorties de classifieurs). Celles que nous avons proposées alors intégraient des informations particulièrement hétérogènes car provenant de niveaux d'abstraction différents (géométrique, topologique, lexical …). L'ensemble a conduit à un puissant moteur de lecture automatique de textes manuscrits. Nous avons montré ensuite, lors de travaux précurseurs dans le domaine, qu'il était possible d'améliorer sensiblement les performances du système de lecture en mettant en place des stratégies d'**adaptation à l'utilisateur** intelligentes, conduisant à spécialiser le système à un scripteur donné.

2002 a constitué une année charnière dans mon parcours avec l'arrêt progressif des travaux en reconnaissance de l'écrit. Arrêt dû, entre autres, au fait qu'une poursuite de ces recherches nous aurait menés vers des domaines très éloignés de nos problématiques habituelles comme l'ergonomie des interfaces ou le traitement du langage naturel. Nous avons initié des travaux en **localisation de visages** avec la thèse de Rachid Belaroussi. Ceci m'a conduit à étudier en détail la phase d'**extraction d'informations dans les données**, précédent la phase de classification dans tout processus de reconnaissance des formes. Glissement aussi vers les méthodes de traitement d'images car le signal à analyser devenait bidimensionnel. Nous avons alors modélisé l'apparence du visage à l'aide de réseaux de neurones auto-associatifs. Utilisés comme détecteurs ou localiseurs de visages dans des images complexes, ces derniers se sont avérés insuffisants. C'est pourquoi nous avons décidé de rechercher d'autres sources d'informations. Les modélisations colorimétrique de la teinte « chair » et géométrique de la forme elliptique du visage, combinées à l'information anthropomorphique précédente, étaient suffisamment décorrélées pour donner au localiseur d'excellentes performances. Une fois de plus, un réseau de neurones a servi pour intégrer l'information. La validation de cette approche passait par sa comparaison avec un algorithme de l'état de l'art : le détecteur de Viola & Jones (une cascade attentionnelle de classifieurs *boostés* que nous décrirons en détail au §2.2).Une analyse approfondie des deux algorithmes a montré leurs limites respectives. Nous avons alors combiné le détecteur de l'état de l'art avec un modèle colorimétrique adaptatif pour réaliser un **suivi du visage** en temps réel, fonctionnant quelque soit la pose du visage. Plus récemment, nous nous sommes penchés sur le problème de **localisation des caractéristiques faciales**. Nous avons proposé une variante originale de réseaux auto-associatifs pour réaliser cette tâche de régression. Une fois de plus, pour gérer ce problème complexe (la localisation doit aboutir quelle que soit l'orientation du visage), nous avons mis en œuvre un ensemble de réseaux qui, dans sa version la plus aboutie, fonctionne en temps réel.

Depuis 2006, à la faveur d'un contrat de recherche avec Peugeot-Citroen Automobiles (PSA) finançant la thèse de Pablo Negri, nous travaillons activement sur le ***boosting***. L'application choisie est la **détection de véhicules** dans des scènes routières. Quoiqu'élégant et apparemment simple à mettre en œuvre, l'algorithme a un comportement très complexe. Il est très utilisé pour les problèmes de détection caractérisés par l'absence d'exemples étiquetés négatifs et la sélection de ces derniers pendant l'apprentissage s'avère cruciale. Il combine les phases d'extraction des caractéristiques et de classification de façon quasi optimale, mais nécessite de trouver l'espace de représentation adéquat. Nous montrons l'importance du choix des caractéristiques et en combinons deux types : générative et discriminante.

En 2007, lors de la création de l'Institut des Systèmes Intelligents et Robotiques (ISIR – CNRS FRE2507), j'ai rejoins l'équipe-projet « lunettes intelligentes » qui vise à créer des dispositifs d'aide à la navigation pour les déficients visuels. La thèse de Shehzad Muhammad Hanif m'a permis d'initier des recherches sur la **localisation de textes** dans les images naturelles. Nous travaillons actuellement sur le processus de sélection de caractéristiques et envisageons de mettre à profit notre expertise sur le boosting pour traiter ce problème.

En 2008 enfin, j'organiserai le workshop du 3ème Comité technique de l'IAPR (*International Association for Pattern Recognition, TC3 : Neural Networks and Computational Intelligence*) : ***Artificial Neural Networks in Pattern Recognition***.

2.2 SYNTHESE

Comme je l'ai dit au départ, les réseaux de neurones ont constitué l'un des fils conducteurs de mes recherches. Mais, le lecteur a pu le constater, le spectre de mes recherches couvre l'ensemble du processus de reconnaissance des formes dont les principales étapes sont représentées dans la figure ci-dessous. A chaque étape, plusieurs choix sont possibles conduisant à un système final différent. Ces différentes réalisations d'une même tâche peuvent être combinées. Plusieurs facteurs sont à l'origine de la coopération : les limites de chaque technique considérée seule, la complémentarité avérée de plusieurs techniques, la puissance de calcul sans cesse croissante des ordinateurs bien sur, mais aussi l'intégration plus aisée dans le silicium … Mais il existe des origines autres que logicielles ou matérielles à la coopération : entomologiques d'abord avec le comportement « intelligent » des colonies de fourmis et des essaims d'abeilles, mais surtout cognitivistes avec le cerveau, fantastique système coopératif ; chaque zone du cortex étant dévolue à une tâche bien précise, et en perpétuelle interaction avec les zones voisines.

Les paragraphes suivants tentent de décrire les différentes étapes du processus de reconnaissance des formes. Ils précisent quelques problèmes-clés et les méthodes les plus couramment mises en œuvre pour résoudre ces derniers.

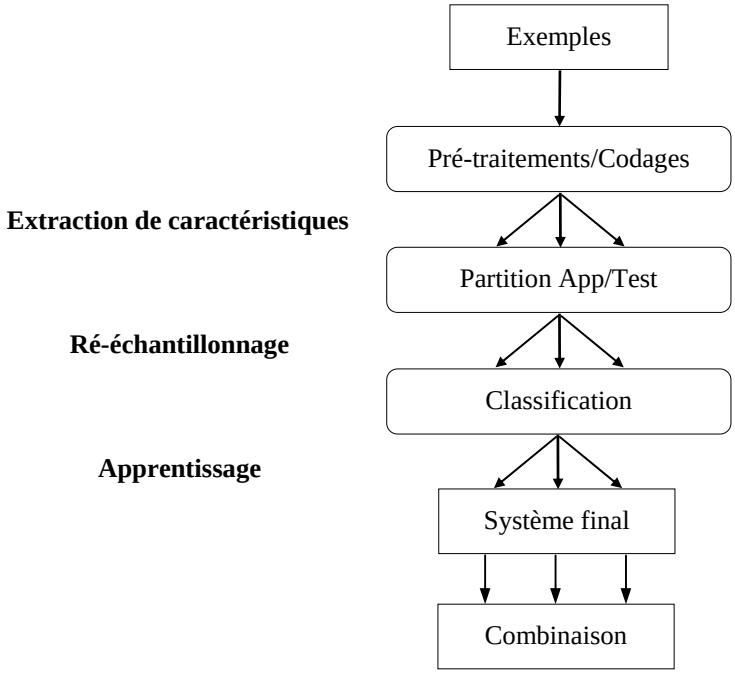


Figure 1 : Les différentes étapes du processus de reconnaissance des formes (phase de paramétrisation).

La majorité des algorithmes décrits dans ce mémoire sont de type supervisé, ce qui suppose l'existence préalable d'une **base d'exemples étiquetés** dédiée à la paramétrisation des systèmes, puis à leur évaluation. Cet étiquetage est une opération fastidieuse, généralement effectuée de façon manuelle ou semi-automatique par un expert. Dans de nombreux domaines toutefois, la communauté est suffisamment organisée pour proposer des *benchmarks*, permettant aux équipes de recherche de travailler sur des données communes.

Les problèmes de reconnaissance des formes se posent généralement en dimension élevée et sont frappés de plein fouet par la malédiction de la dimensionnalité. La phase d'**extraction de caractéristiques** s'avère donc cruciale L'exemple classique est l'image constituée de pixels en très grand nombre. La quantité d'information véhiculée par chaque pixel est faible et très redondante. Il est préférable d'extraire de l'image des informations de plus haut niveau (contours, régions …). A l'issue de l'étape de génération des caractéristiques, la dimension des données peut être très importante, d'où la nécessité d'extraire de celles-ci les variables les plus pertinentes. Parmi les méthodes d'extraction de caractéristiques, citons en premier la moins satisfaisante : celle réalisée par l'expert et guidée par les données, ou par la tâche à accomplir … qui présente toutefois l'avantage d'être interprétable puisque basée sur des caractéristiques objectives. L'extraction peut se faire automatiquement par sélection des caractéristiques ou par génération de nouvelles caractéristiques, combinaisons des premières. Parmi les algorithmes de sélection, on différencie les méthodes de filtrage (séquentielles [AD91[2]] ou génétiques [OSB03[3]]) utilisant une fonction d'évaluation (*fitness*) des méthodes « *wrapper* » où la fonction d'évaluation est le résultat de classification. Les méthodes séquentielles par ajout (*Forward Selection*) et par suppression (*Backward Deletion*) ont été combinées dans un algorithme générant des solutions quasi optimales : *Floating Search* [PNC94[4]]. Issues de l'analyse statistique, les méthodes de génération par projection dans un sous-espace sont largement utilisées : non supervisées comme l'analyse en composante principales ou indépendantes (dont il existe des versions supervisées) ; supervisées comme l'analyse discriminante. Le propos n'étant pas de détailler l'ensemble de ces méthodes, nous invitons le lecteur à consulter l'excellent ouvrage [GGN06[5]].

L'étape de partition en bases d'estimation (apprentissage) et d'évaluation (test) fait appel aux méthodes de **ré-échantillonnage de données**. La méthode *hold out* et ses déclinaisons sont couramment mises en œuvre : *leave one out* lorsque l'échantillon est de petite taille (solution peu satisfaisante car produisant une estimation de variance forte) ; *leave many out*, plus satisfaisant, dans le cas général. La synthèse d'exemples, en augmentant la taille de l'échantillon et le *bootstrap*, en modifiant la distribution statistique des données, permettent d'améliorer les capacités de généralisation du système, tout comme le *bagging* (*Bootstrap AGGregatING*) [B96[6]] et le *boosting* (« dopage ») [FS97[7]], que nous décrirons en détail à la fin de ce paragraphe. Notons que ces méthodes peuvent être analysées sous l'angle de la fusion d'information dès lors qu'elles combinent plusieurs réalisations d'un même classifieur.

Les méthodes de **classification** présentées dans ce mémoire sont de type supervisé, nous avons donc pris le parti de ne pas évoquer la découverte de classe (classification non supervisée). Les nombreuses méthodes d'apprentissage supervisé se rattachent à la famille des méthodes d'apprentissage par induction [CMK02[8]]. Toutes adaptent (apprennent, estiment) leurs paramètres en vue de séparer des points (exemples) de l'espace des caractéristiques et

[2] Almuallim H. & Dietterich T.G., Learning with many irrelevant features, National Conference on Artificial Intelligence, pp 547-552, 1991,

[3] Oliveira, L.S., Sabourin, R., Bortolozzi, F. et Suen, C.Y., A Methodology for Feature Selection Using Multi-Objective Genetic Algorithms for Handwritten Digit String Recognition, the International Journal of Pattern Recognition and Artificial Intelligence, 17(6), pp 903-929, 2003.

[4] Pudil P., Novovicova J. & Kittler J., Floating search methods in feature selection, Pattern Recognition Letters, 15(11) : 1119-1125, 994.

[5] Feature Extraction, Foundations and Applications, Guyon I, Gunn S., Nikravesh M. & Zadeh L., Editors, Series Studies in Fuzziness and Soft Computing, Physica-Verlag, Springer, 2006.

[6] Breiman L., Bagging predictors, Machine Learning, 24(2) : 123-140, 1996.

[7] Freund Y & Schapire R.E., A decision-theoretic generalization of on-line learning and an application to boosting, Journal of Computer System Sciences, 55 : 119-139, 1997.

[8] Cornujéols A., Miclet L. & Kodratoff Y., Apprentissage artificiel, Eyrolles, 2002.

donc de distinguer les différentes classes. Plusieurs taxinomies sont proposées dans la littérature. Certains auteurs distinguent les méthodes génératives (qui modélisent les classes) des méthodes discriminantes (qui modélisent des frontières entre les classes) ; d'autres, les méthodes paramétriques des méthodes non paramétriques … Nous différencierons aussi les méthodes de classification « pures » où les exemples de toutes les classes sont étiquetés des méthodes de **détection**, binaires, où seuls des exemples de la classe (positifs) sont disponibles et où les exemples de la « non-classe » (négatifs) doivent être choisis judicieusement c'est-à-dire proches de la frontière et susceptibles d'être classés positifs. Sans prétendre à l'exhaustivité, on citera les méthodes statistiques paramétriques (mélange de distributions, le plus souvent gaussiennes) ou non (noyaux de Parzen, *k*-plus proches voisins [M93[9]]), les arbres de décisions [SL91[10]] et les réseaux de neurones [B95[11]]. Ces derniers ont connu de multiples déclinaisons depuis le perceptron multi-couches initial : réseaux à fonctions radiales (dérivés des modèles multi-gaussiens), réseaux à convolution (qui intègrent la phase d'extraction de caractéristiques dans l'apprentissage [WHH89[12]]) et plus récemment, issus des travaux théoriques de Vapnik [V95[13]], machines à vecteurs supports (*Support Vector Machines* ou SVM). Ces derniers permettent de trouver des hyper-surfaces discriminantes quasi optimales. Ils projettent les données dans un espace de re-description de très grande dimension (projection équivalente à une génération de caractéristiques) où les données sont supposées linéairement séparables et s'appuient sur un nombre limité de vecteurs supports pour maximiser la marge. La comparaison entre méthodes statistiques et neuronales ne permet pas de conclure quant à la supériorité d'une approche sur l'autre [GRB00[14]]. Toutes deux doivent faire face au problème du contrôle des paramètres d'apprentissage et de la complexité du modèle. Sur ce point, le rasoir d'Occam (principe du minimum de complexité) n'a pas été remis en cause, mais des méthodes de régularisation ont vu le jour dans le cadre statistique (critère d'Akaike [A94[15]]) comme dans le cadre neuronal (constructive comme *StepNet* [K92[16]] ou par élagage comme *Optimal Brain Damage* [LDS90[17]]).

Comme il est difficile, voire impossible, de sélectionner à la fois le meilleur ensemble de caractéristiques et l'architecture de classification adéquate en ne disposant que d'un échantillon fini de données, les algorithmes quels qu'ils soient convergeront la plupart du temps vers une solution particulière, inefficace dans certaines situations. La **combinaison** de plusieurs réalisations peut se révéler plus efficace, ce qui motive les nombreuses recherches dans le domaine de la fusion d'informations et des systèmes multi-classifieurs. Ces derniers obéissent au principe « diviser pour mieux régner » et décomposent le problème en sous tâches plus aisément réalisables, en vue d'améliorer les performances, la fiabilité, la robustesse et l'intelligibilité du processus de reconnaissance des formes. De très nombreuses stratégies de combinaison ont vu le jour durant cette dernière décennie. La plupart reposent

[9] Milgram M., Reconnaissance des formes. Méthodes numériques et connexionnistes, Armand Colin, 1993.
[10] Safavian S.R. & Landgrebe D., A survey of decision tree classifier methodology, IEEE Transactions on Systems, Man and Cybernetics, 21(3): 660-674, 1991.
[11] Bishop C. M., Neural Networks for Pattern Recognition, Oxford University Press, 1995.
[12] Waibel A., Hanazawa T., Hinton G., Shikano K. & Lang K., Phoneme recognition using time-delay neural networks, IEEE Transactions on Acoustic Speech &Signal Processing, 37(3), pp 328-339, 1989.
[13] Vapnik V.N., The nature of statistical learning theory, Springer-Verlag, 1995.
[14] Giacinto G., Roli F. & Bruzzone L., Combination of neural and statistical algorithms for supervised classification of remote-sensing images, Pattern Recognition Letters, 21 : 385-397, 2000.
[15] Akaike H., A new look at statistical model identification, IEEE Transactions on Automatic Control, 19: 716-723, 1994.
[16] Knerr S., Personnaz L. & Dreyfus G., Une nouvelle approche de la reconnaissance de chiffres manuscrits par réseaux de neurones, Congrès National sur l'Ecrit et le Document, pp 325-332, 1992.
[17] LeCun Y., Denker J., Solla S., Howard R. E. & Jackel L. D., Optimal brain damage, Advances in Neural Information Processing Systems II, 1990.

sur des heuristiques, mais plusieurs contributions récentes tentent de mieux formaliser le problème et d'établir un cadre théorique commun [K98[18], KW03[19], FR05[20]]. Toutes montrent l'importance de la diversité des classifieurs individuels composant le système. Il n'existe pas de règles formelles de construction d'un ensemble de classifieurs. Toutefois, certains auteurs suivent le paradigme « sur-production et sélection », qui préconise de générer un grand nombre de classifieurs, puis de sélectionner et de combiner les plus efficaces [RGV01[21]]. Une fois générés ces classifieurs, le concepteur doit encore choisir la règle de combinaison : méthodes de vote [XKS92[22]], règles « somme » et « produit », réseaux de neurones [HS94[23]], méthodes probabilistes, possibilistes ou basées sur la théorie de l'évidence [B03[24]].

Nous terminons ce bref exposé en présentant les techniques de ***boosting***. L'algorithme AdaBoost, proposé initialement dans [FS97], s'est montré capable d'améliorer les performances de nombreux systèmes de classification et de détection. Il trouve une hypothèse précise *G* (appelée classifieur fort) en combinant plusieurs fonctions de classification faibles g_t qui, en moyenne, ont une précision modérée (en fait, légèrement meilleures que le hasard). La fonction de décision de l'algorithme (*discrete*) AdaBoost évalue la somme pondérée (par les coefficients α_t) des sorties des classifieurs faibles :

$$G = \begin{cases} 1 & \sum_{t=1}^{T} \alpha_t g_t \geq \frac{1}{2} \sum_{t=1}^{T} \alpha_t = S \\ 0 & \text{sinon} \end{cases}$$

Le paramètre *T*, qui correspond au nombre de classifieurs faibles pris en compte dans la combinaison, est réglé pendant l'apprentissage, pour atteindre un objectif fixé par l'utilisateur (typiquement, un taux de reconnaissance, de détection ou de fausses alarmes). Les hypothèses de classification faibles g_t sont apprises itérativement, sur des distributions de données différentes, obtenues en surpondérant au fur et à mesure les exemples mal classées aux étapes précédentes. En fin d'apprentissage, l'algorithme se concentre sur les exemples les plus difficiles à classer (ce qui peut perturber la convergence). Les coefficients α_t sont inversement proportionnels à l'erreur faite par les classifieurs faibles g_t et donnent un poids plus important aux meilleures hypothèses. L'analyse théorique de l'algorithme montre qu'il aurait tendance à sur-apprendre pour de grandes valeurs de *T*. L'analyse empirique contredit cette affirmation. Il apparaît même fréquemment que l'erreur de test continue de diminuer alors que l'erreur d'apprentissage est nulle. La justification théorique de ce comportement tient au fait qu'AdaBoost, comme les SVM, recherche des fonctions de classification à marge maximale [SFB97[25]]. Il est intéressant de revisiter cet algorithme. Du point de vue de la fusion de données, on retrouve la règle de décision par somme pondérée seuillée. Vu sous l'angle des réseaux de neurones, ensuite, on voit apparaître un réseau mono-couche linéaire à seuil dont les connexions α_t, mais aussi le nombre d'entrées *T*, sont adaptés pendant l'apprentissage.

[18] Kittler J., Combining Classifiers: A theorical Framework, Pattern Analysis & Applications, 1, pp 18-27, 1998.
[19] Kuncheva L.I. & Whitaker C.J., Measures of diversity in classifier ensembles, Machine Learning, 51, 181-207, 2003.
[20] Fumera G.; Roli F., A theoretical and experimental analysis of linear combiners for multiple classifier systems, IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(6), pp 942-956, 2005.
[21] Roli F., Giacinto G. & Vernazza G., Methods for Designing Multiple Classifier Systems, Multiple Classifier Systems, LNCS 2096, pp 78–87, 2001.
[22] Xu L., Kryzak A. & Suen C.Y., Method of combining multiple classifiers and their application to handwriting recognition, IEEE Transactions on Systems, Man & Cybernetics, 22(3), pp 418-435, 1992.
[23] Huang Y.S. & Suen C.Y., A method of combining multiple classifiers - a neural network approach, International Conference on Pattern Recognition, (B), pp 473-475 1994.
[24] Bloch I., Fusion d'informations en traitement du signal et des images, Lavoisier, 2003.
[25] Schapire R.E., freund Y., Bartlett P. & Lee W.S., Boosting the margin: a new explanation for the effectiveness of voting methods, International Conference on Machine Learning, pp 322-330, 1997.

L'algorithme proposé dans [VJ01[26]] est une transposition à l'image d'AdaBoost. Il comporte deux innovations majeures. En premier lieu, il utilise comme fonctions de classification faibles les sorties de filtres rectangulaires (de Haar) de taille variable qui, lorsqu'ils sont appliqués à une image, permettent d'extraire les contours des objets. Le nombre de classifieurs faibles obtenus en déplaçant les filtres sur l'ensemble de l'image et en modifiant leur taille est très grand. La procédure de sélection précédente permet (suivant le paradigme « *overproduce and select* ») de trouver les *T* caractéristiques (sorties des filtres) les plus discriminantes. La principale innovation est la mise en place d'une **cascade attentionnelle** qui combine séquentiellement *N* classifieurs forts afin de remplir un objectif multi-critères (taux de détection maximum ET taux de fausses alarmes minimum dans un cadre de détection). Une donnée doit donc traverser tous les étages de la cascade pour être validée. L'algorithme initial (un seul classifieur fort) nécessitait un grand nombre de caractéristiques (*T*) pour atteindre l'objectif. La cascade attentionnelle permet de diminuer drastiquement ce nombre, diminuant d'autant le temps de calcul, tout en améliorant les performances du système. On voit ici toute l'élégance de la méthode, qui combine astucieusement les phases d'extraction de caractéristiques, de multi-classification et de fusion d'informations.

2.3 ORGANISATION DU MEMOIRE

J'ai pris le parti de présenter mes contributions chronologiquement (figure 2) et non thématiquement, afin de donner au document une plus grande clarté. C'est pourquoi je commencerai cet exposé par mes recherches doctorales et post-doctorales en classification de caractères isolés (§3.1). La section suivante sera dédiée aux travaux réalisés, pendant la thèse de Loïc Oudot, sur la lecture automatique de textes manuscrits et l'adaptation à l'utilisateur (§3.2). J'exposerai ensuite les recherches effectuées en analyse de visages, dont une grande partie correspond à la thèse de Rachid Belaroussi (§3.3). Dans la section 4, je présenterai les travaux que nous menons actuellement avec Pablo Negri, en détection de véhicules (§4.1), avec Shehzad Muhammad Hanif, en localisation de textes (§4.2). Les conclusions et perspectives de la section 5 seront suivies d'un recueil d'articles et de communications permettant d'approfondir cette présentation.

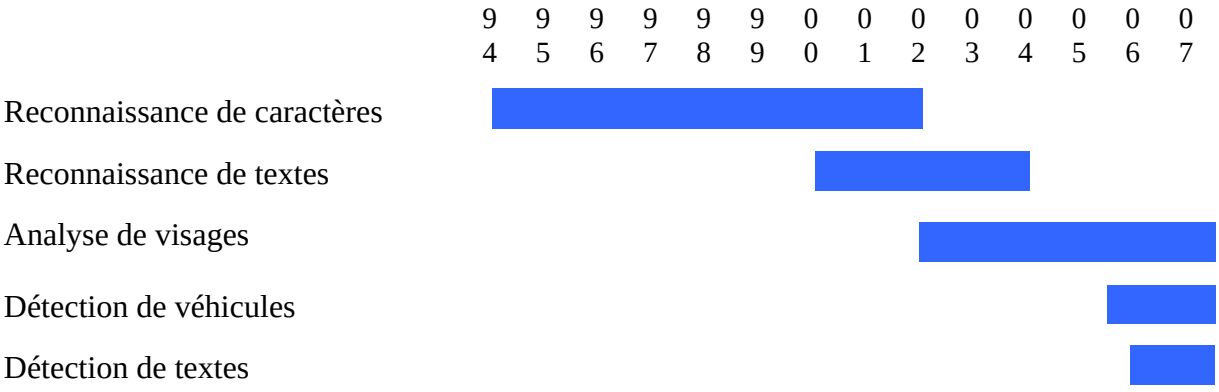


Figure 2 : Chronologie des activités de recherche.

[26] Viola P. & Jones M., Rapid object detection using a boosted cascade of simple features, International Conference on Computer Vision and Pattern Recognition, (1), pp 511-518, 2001.

3. CONTRIBUTIONS

3.1 CLASSIFICATION DE CARACTERES ISOLES

3.1.1 Contexte

Lorsque ces recherches ont débuté en 1994, le Newton, premier assistant personnel développé par Apple, connaissait un échec commercial. Il faut dire que le concept du "papier électronique" [HF91[27], P91[28]] était particulièrement audacieux, puisqu'il rassemblait en un cocktail explosif les avantages du multi-fenêtrage, de la reconnaissance de l'écrit et de la parole, du multimédia ... En regard de quoi le Newton, avec son système de reconnaissance laborieux, ne pouvait paraître que faiblement attractif. Malgré cela, la communauté DAR (*Document Analysis and Recognition*) restait très active, particulièrement en France. C'est dans le domaine de la reconnaissance statique (traitement hors ligne de documents scannés) qu'était conduite la majorité des recherches, motivées par des enjeux économiques : lecture automatique du code puis de l'adresse postale, du montant de chèque, des champs dans les formulaires … La reconnaissance dynamique (traitement en ligne de la trace d'un stylo électronique sur une tablette à digitaliser ou un écran tactile) était aussi très active.

L'idée directrice du projet était de réaliser un système de reconnaissance et d'analyse des expressions mathématiques manuscrites, en vue de remplacer, pour toutes les opérations de saisie et d'édition d'équations et de formules mathématiques, chimiques … le clavier et la souris par un stylo, plus convivial. Ce projet nécessitait la mise au point de systèmes de reconnaissance d'autant plus performants qu'ils seraient confrontés à une série de problèmes neufs qui n'apparaissaient pas en reconnaissance d'écriture « classique », fut-elle cursive. Citons en particulier l'aspect bidimensionnel caractéristique des objets mathématiques, l'absence de contraintes lexicales (matérialisées par un dictionnaire) permettant d'optimiser les performances en corrigeant les erreurs de classification et surtout, la taille de l'alphabet incluant les chiffres, majuscules et minuscules latin, les lettres grecques et de nombreux symboles (au lieu des 26 minuscules de l'alphabet classique). L'analyse de l'existant faisait surtout état de systèmes mono-scripteur (dédiés à la reconnaissance de l'écriture d'une personne en particulier). Nous souhaitions nous affranchir de cette contrainte et travailler dans un cadre omni-scripteur (« universel »). Face à la complexité de ce problème quasiment non contraint, nos recherches se sont concentrées sur l'étape de classification de caractères.

Deux types de méthodes s'opposent dans ce domaine : les classifieurs générateurs et les classifieurs discriminants. Les premiers estiment les paramètres d'un (ou plusieurs) modèle(s) pour chaque classe de caractère en se basant uniquement sur les exemples de cette classe. En phase de test, la classification est effectuée *via* une mesure de similarité (ou de distance) entre la donnée à classer et les modèles, suivie d'une mise en compétition. Les modèles neuronaux [SM96[29]], markoviens [CJ02[30]], hybrides [G96[31], PVL04[32]], flous [AL96[33]], possibilistes

[27] Higgins C.A. & Ford D.M., Stylus driven interfaces - the electronic paper concept, International Conference on Document Analysis and Recognition, (B), pp 853-862, 1991.
[28] Plamondon R., Step toward the production of an electronic pen-pad, International Conference on Document Analysis and Recognition, (A), pp 361-371, 1991.
[29] Schwenk H. & Milgram M. Constraint tangent distance for on-line character recognition, International Conference on Pattern Recognition, (D), pp 520-524, 1996.
[30] Connell S. & Jain A.K. Writer adaptation of on-line handwriting models, IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(2): 329-342, 2002.

[QDM05[34]] ou à prototypes [VLK02[35]] sont les plus couramment utilisés. Tous ces modèles sont parfaitement adaptés au contexte omni-scripteur. Les classifieurs discriminants déterminent des frontières optimales entre les classes et sont donc entraînés à l'aide d'exemples de plusieurs classes. Les réseaux à connexions locales et poids partagés décrits dans [GAL91[36]] séparent toutes les classes. Dans [OS02[37]], N réseaux sont entraînés pour discriminer une classe de toutes les autres. Cette stratégie « un contre tous » s'avère plus efficace que la précédente. Une autre stratégie (« un contre un ») consiste à entraîner un classifieur pour chaque paire de classes [PKP94[38]]. Enfin, la combinaison des deux approches (générative et discriminante) permet d'augmenter l'efficacité globale [RSN03[39]]. Plus généralement, la reconnaissance de l'écriture manuscrite est un problème d'une telle complexité (en raison de la forte variabilité du signal traité) que la combinaison de différents clasifieurs entraînés sur des données ou des caractéristiques différentes est une tendance forte des recherches menées actuellement [JDM00[40], RF03[41]].

Nous présentons dans la suite plusieurs travaux dédiés à la reconnaissance de caractères isolés. Toutes les expérimentations ont été conduites sur la base UNIPEN *Train R01-V07* [GSP94[42]] qui compte plus de 80 000 exemples de caractères. Elle a été divisée en trois corpus : d'estimation B_{ES}, de validation croisée B_{VC} et de test B_{TE}. Les pré-traitements sont réduits : la séquence des coordonnées *(x,y)* formant le caractère est ré-échantillonnée à raison de 20 points par traits, centrée et normalisée en conservant le ratio hauteur/largeur.

3.1.2 Modélisation non supervisée des allographes de caractères [Annexe 1]

Aux impératifs de l'application précédemment cités (alphabet étendu et contexte omni-scripteur) s'ajoutait une contrainte d'incrémentalité permettant, d'une part, l'adjonction de nouvelles classes de symboles et, d'autre part, l'activation et l'inhibition aisée de groupes de classes en vue d'accélérer la phase de classification. Seul un classifieur générateur, entraîné sur les seuls exemples d'une classe et modélisant celle-ci par un (ou plusieurs) expert(s), pouvait satisfaire cette contrainte. Le contexte omni-scripteur entraînant la multiplication des allographes (variantes) de caractères, écartait d'emblée certaines techniques classiques du

---

[31] Garcia-Salicetti S., Une approche neuronale prédictive pour la reconnaissance en-ligne de l'écriture cursive, Thèse de Doctorat de l'Université Paris VI, 1996.

[32] Poisson E., Viard-Gaudin C. & Lallican P.M., Système TDNN/HMM de reconnaissance de mots cursifs en ligne à apprentissage simplifié, Colloque International Francophone sur l'Ecrit et le Document, 2004.

[33] Anquetil E. & Lorette G. On-line Handwriting Recognition system Based on Hierarchical Qualitative Fuzzy Modeling, International Workshop on Frontier of Handwriting Recognition , pp 47-52, 1996.

[34] Quost B., Denoeux T. & Masson M., Pairwise classifiers in the framework if belief functions, International Conference on Information Fusion, 2005.

[35] Vuori V., Laaksonen J. & Kangas,J. Influence of erroneous learning samples on adaptation in on-line handwriting recognition, Pattern Recognition, 35(4): 915-925, 2002.

[36] Guyon I., Albrecht P., LeCun Y., Denker J. & Hubbard W. Design of a Neural Network Character Recognizer for a Touch Terminal, Pattern Recognition, 24(2): 105-119, 1991.

[37] Oh I.S. & Suen C.Y.: A class-modular feed-forward neural network for handwriting recognition, Pattern Recognition, 35: 229-244, 2002.

[38] Price D., Knerr S., Personnaz L. & Dreyfus G. Pairwise neural network classifiers with probabilistic outputs, Neural Information Processing Systems, 7, 1994.

[39] Raina R., Shen Y., Ng A.Y. and McCallum A. Classification with hybrid generative/discriminative models, Neural Information Processing Systems 16, 2003.

[40] Jain A.K., Duin R. & Mao J. Statistical pattern recognition: a review, IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(1): 4-37, 2000.

[41] Rahman A.F. & Fairhust M.C. Multiple classifier decision combination strategies for character recognition: a review, International Journal on Document Analysis and Recognition, 5: 166-194, 2003.

[42] Guyon I., Schomaker L., Plamondon R., Liberman M. & Janet S. UNIPEN project of on-line data exchange and recognizer benchmarks, International Conference on Pattern Recognition, pp. 29-33, 1994.

domaine (neuronales, markoviennes ou floues) en raison du faible volume de données disponibles pour entraîner les modèles et rendait idéale la conception d'un système à génération de prototypes. Notons qu'une solution élégante au précédent problème de l'absence de données a été proposée dans la suite (§2.1.4) et dans [MA06[43]] : la synthèse d'exemples par transformation. J'ai donc développé un algorithme de *clustering* original permettant d'estimer l'ensemble des allographes pertinents d'une classe de symbole. Baptisé SMAC, il est basé sur une modélisation intrinsèque de chaque classe, ne nécessite aucune initialisation et déterminer le nombre et la nature des prototypes nécessaires pour représenter correctement la classe. Il commence par une phase de séparation morphologique (SM) des caractères en sous-classes homogènes en fonction du nombre de *strokes* (portion du signal écrit compris entre deux levers de stylo) composant le tracé. Suivent des phases d'agglomération compétitives (AC), inspirées des techniques de classification ascendantes hiérarchiques [RLE98[44]] qui permettent de trouver les prototypes pertinents dans chaque sous-classe. La classification se fait ensuite par l'algorithme du plus proche voisin, qui utilise comme métrique la distance élastique [SMP96[45]]. Les performances obtenues sont inférieures à celles du classifieur direct (sans compilation des références : Tableau 1). Le taux de reconnaissance atteint toutefois 98% sur les chiffres et 95.5% en moyenne sur les 80 classes de caractères ; et la vitesse de classification a été multipliée par 20.

Algorithme	**1-ppv**	**SMAC**	**LVQ**	**SMAC2**
Taux d'erreur	0.7%	2.1%	1.8%	**1.2%**
Mémoire	2320 Ko	94 Ko	94 Ko	**94 Ko**
Vitesse relative	1	22	22	**22**

Tableau 1. Taux d'erreur des différents algorithmes sur les chiffres (base de test : B_T).

Une étude approfondie de l'algorithme et sa comparaison avec l'algorithme de quantification vectorielle LVQ [K89[46]], ont montré les qualités et les défauts des deux méthodes. Un nouvel algorithme hybride (baptisé SMAC2) a alors été proposé. Il utilise l'algorithme SMAC pour initialiser l'algorithme LVQ. Le nombre optimal de prototypes par classes est ainsi déterminé. Ces prototypes sont ensuite modifiés par quantification vectorielle, qui optimise leur représentativité en minimisant la variance intra-classe. Le taux de reconnaissance atteint alors 99% sur les chiffres et près de 97%, en moyenne, sur les 80 classes. L'analyse des erreurs de classification (Figure 1) permet d'émettre deux hypothèses quant aux origines de l'échec du classifieur dynamique face à certains exemples :

- Des exemples dont l'image (donc la représentation statique) ne peut porter à confusion ... et qui sont pourtant confondus. Pour tous ces allographes, aucun prototype (de leur classe) n'a été trouvé. La dynamique utilisée est, soit fortement confusive (à l'image des "9" écrits comme des "5"), soit inhabituelle (quelques "4" et "7" en particulier). Pour tous ces exemples, il semble possible de lever l'ambiguïté en utilisant

[43] Mouchère H. & Anquetil E., Synthèse de caractères manuscrits en-ligne pour la reconnaissance de l'écriture, Colloque International Francophone sur l'Ecrit et le Document, pp 187-192, 2006.
[44] Ribert A., Lecourtier Y., Ennaji A. & Stocker E., Vers un classifieur neuronal incrémental : une construction évolutive de taxinomies numériques, Colloque International Francophone sur l'Écrit et le Document, pp 141-150, 1998.
[45] Schwenk H., Prevost L. & Milgram M., Comparaison de la distance élastique et de la distance tangente en reconnaissance de caractères on-line, RFIA'96 (Reconnaissance des Formes et Intelligence Artificielle), (2), pp 589-596, Rennes, 1996.
[46] Kohonen T., Self-organisation and associative memory, Springer-Verlag, 3rd Ed, 1989.

conjointement les représentations dynamique et statique des données, coopération que nous présenterons dans le prochain paragraphe.

- Des exemples particulièrement ambigus (ou même l'expert humain ne pourrait se prononcer) : "1" ressemblant à un "7", "0" écrits comme un "6",... Dans ce cas, on atteint les limites des classifieurs générateurs : les deux classes de caractères étant entremêlées, les frontières grossières définies par ces classifieurs sont incorrectes. Cette fois, l'ambiguïté peut être levée à l'aide de classifieurs discriminants locaux, entraînés à raffiner la frontière.

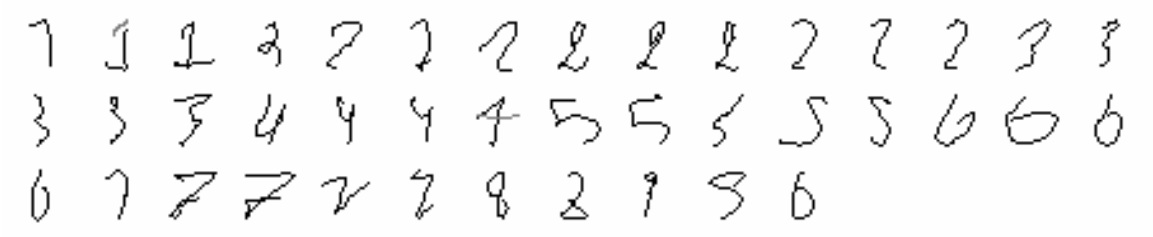


Figure 1. Erreurs de classification (base de test : B_T).

3.1.3 Fusion d'informations statique et dynamique

J'ai développé un second classifieur, basé sur le même principe que le précédent (clustering des références et classification par l'algorithme du plus proche voisin), mais traitant les données statiques (caractères dynamiques transformés en images matricielles) et utilisant comme métrique la distance de Mahalanobis. Malgré des performances bien inférieures (90% sur les chiffres et 86% sur les 80 classes), ce classifieur présentait l'avantage d'être suffisamment indépendant du précédent pour produire des décisions décorrélées. Des mesures simples (assez proches de celles proposées dans [KW03[47]]), comme le nombre d'exemples pour lesquels les deux classifieurs entrent en conflit (environ 10% des données) ou proposent tous les deux une même réponse erronée (0.3%), ont permis de le vérifier. Faire coopérer ces deux classifieurs paraissait donc prometteur.

3.1.3.1 Architecture parallèle

Le principe de base de cette stratégie est de présenter simultanément la forme à reconnaître à l'ensemble des classifieurs. Les sorties de ces derniers sont ensuite normalisées, puis concaténées pour former le vecteur d'entrée de l'expert de fusion qui prend la décision finale.
L'inconvénient majeure d'une telle approche est qu'elle nécessite d'activer l'ensemble des modules du système. Par contre, la décision finale est prise avec le maximum de connaissances mises à disposition par chaque expert. Deux techniques de fusion paramétrique ont été envisagées : la moyenne pondérée (stratégie PM) des sorties et un réseau de neurone intégrateur (stratégie PNI). La première, qui pondère indifféremment toutes les sorties d'un classifieur apporte une amélioration relative de 25%. La seconde réalise une pondération plus fine et non linéaire qui porte l'amélioration relative à 40%

[47] Kuncheva L.I. & Whitaker C.J., Measures of diversity in classifier ensembles, Machine Learning, 51, 181-207, 2003.

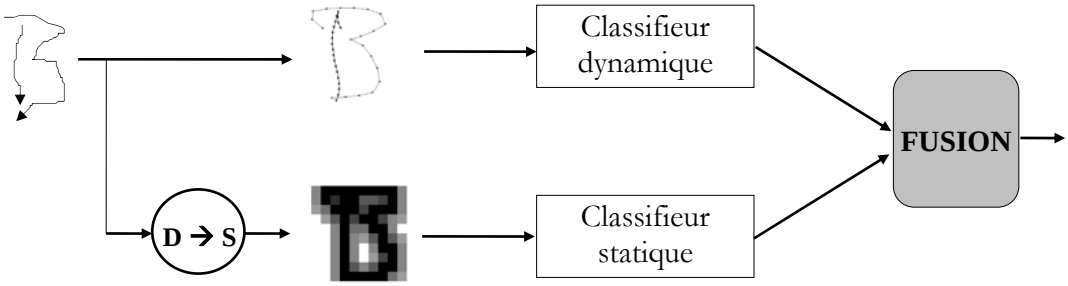


Figure 2. Combinaison parallèle.

3.1.3.2 Architecture Cascade

Il s'agit d'un croisement entre une architecture séquentielle (où les sorties d'un module de traitement servent d'entrées au module suivant), une architecture sélective (où chaque module oriente l'information vers un autre expert plus spécialisé) et une architecture intégratrice. Cette méthode doit permettre de contourner les inconvénients de l'architecture parallèle évoqués précédemment. Ici, le premier classifieur va être suivi par un module de sélection qui va jouer un rôle discriminant en sélectionnant les *K* meilleures hypothèses. Cette sélection sera réalisée en classant les scores réalisés par le premier module pour ne conserver que les *K* meilleurs. Seuls les *K* sous-modules correspondants du classifieur suivant seront activés (Figure 3). En associant cette architecture et la fusion par somme pondérée (stratégie CM), on retrouve les performances de la stratégie PM mais la vitesse de classification a augmenté spectaculairement grâce à la faible valeur de *K*.

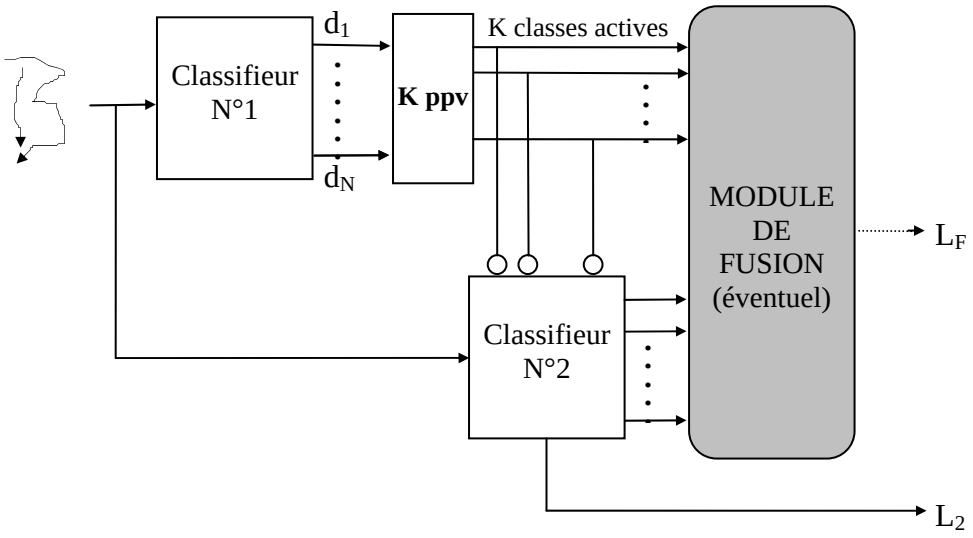


Figure 3. Combinaison hybride (série/parallèle).

3.1.4 Fusion de classifieurs générateurs et discriminants [Annexe 2]

Encadrement : 1 projet (Michel-Sendis C., Moises A.)

3.1.4.1 Principe de la cascade

Comme on l'a vu précédemment, certains caractères manuscrits sont tellement ambigus qu'un classifieur générateur sera incapable de les identifier. De fait, la mise en compétition de deux modèles de classes est inefficace si ces dernières sont fortement entrelacées. Dans ce cas toutefois, le problème se réduit à déterminer la plus pertinente des deux premières hypothèses proposées par le classifieur générateur. Les résultats de classification (Tableau 2) montrent que considérer non pas la meilleure réponse (Top1), mais les deux meilleures (Top2), augmente sensiblement le taux de reconnaissance. Reste dès lors à trouver une méthode susceptible de traiter localement les confusions apparaissant entre ces deux classes. Un simple MLP discriminant bi-classes (*pairwise neural net* PNN) permet de réaliser cette tâche.
On suppose d'une part que le comportement du classifieur est décrit par sa matrice de confusion ; d'autre part que son comportement *a priori* (sur la base d'estimation) est représentatif de son comportement *a posteriori* (sur la base de test). Dès lors, les couples de classes confusives *(i,j)* peuvent être trouvés en analysant la matrice de confusion. Cette dernière (Figure 4) est déterminée sur l'ensemble d'estimation B_{ES}.

	Classifieur générateur		**Classifieur hybride**
	Top1	Top2	
Chiffres	98.9 %	99.8 %	**99.1 %**
Majuscules	96.7 %	99.0 %	**97.9 %**
Minuscules	96.3 %	98.8 %	**97.8 %**

Tableau 2 : Taux de reconnaissance du classifieur générateur et du classifieur hybride (base de test : $B_{VC} \cup B_{TE}$).

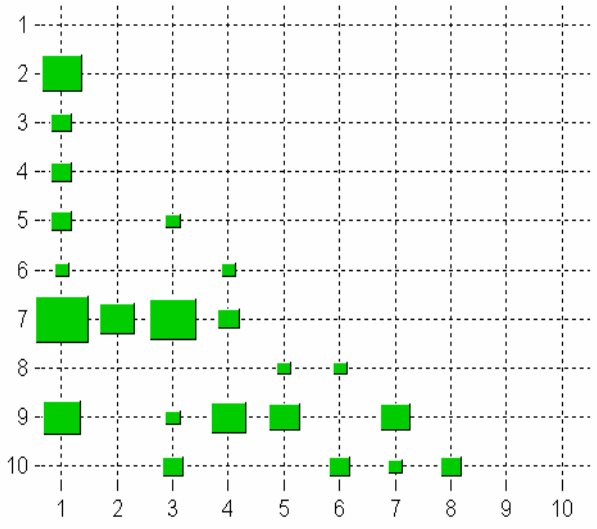


Figure 4 : Matrice de confusion (chiffres : base d'estimation B_{ES}).

Trois conclusions s'imposent lorsqu'on examine la matrice de confusion :

- De nombreux couples ne produisent aucune confusions : (4,2), (5,2) ... il est donc inutile d'entraîner un PNN pour ces couples.
- Certains couples produisent quelques confusions, il peut être utile d'entraîner un PNN.
- Quelques couples sont à l'origine de la majorité des confusions : (1,2), (1,7) …Ces couples doivent impérativement être discriminés.

On peut donc choisir de ne traiter que les couples dont la probabilité de confusion (déterminée sur la base d'estimation) dépasse un certain seuil δ_{conf} et régler ainsi la robustesse du classifieur hybride et le nombre de PNN créés. Pour chaque paire ambiguë, un PNN est entraîné. Il s'agit dès lors d'un problème de classification binaire : la classe *i* est associée au label +1 et la classe *j* au label –1. Les bases d'apprentissage des classes i et j n'ont généralement pas la même taille. Cette dernière est directement proportionnelle à la fréquence d'apparition des lettres de classes i et j dans le langage écrit. De plus, le nombre d'exemples confondus est généralement très faible. La frontière entre les deux classes n'est pas bien définie et le PNN correspondant généralise mal. Pour surmonter cet inconvénient, les exemples confondus de la base d'apprentissage ont été détectés et leur nombre augmenté artificiellement par transformations (rotation, dilatation et contraction). On retrouve là un des principes directeurs du *boosting* : la densité de probabilité des exemples confondus, proches de frontière augmente pendant l'apprentissage.

Les expériences conduites pour différentes valeurs de δ_{conf} montrent logiquement l'augmentation du nombre de PNN crées lorsque la probabilité de prise en compte de la confusion diminue. Toutefois, pour un seuil de confusion minimal, seuls 112 PNN ont été entraînés (contre 695 en entraînant des PNN suivant une stratégie *one against one*, en considérant séparément les chiffres, les majuscules et les minuscules). Ceci est dû au fait que de nombreuses paires de classes ne sont pas confusives.

En phase d'utilisation, pour une forme inconnue de l'ensemble de test, le classifieur générateur fournit le couple de classes le plus pertinent.

- Si le couple (C_1,C_2) n'appartient pas à l'ensemble des classes couramment confondues, on conserve la réponse du classifieur par modélisation (C_1).
- Sinon, le PNN est activé afin de discriminer les deux hypothèses.

Le classifieur hybride a été évalué sur la base de test. On peut constater (Tableau 2) que ses performances sont sensiblement meilleures que celles du classifieur générateur initial.

3.1.4.2 Cascade incrémentale

Dans cette section, nous montrons que le classifieur hybride précédemment décrit est parfaitement adapté à l'apprentissage incrémental c'est à dire que l'adjonction de nouvelles classes peut se faire de façon aisée (sous réserve de conserver en mémoire la base d'apprentissage/estimation initiale). L'étude qui suit est juste une simulation : nous considérons que le classifieur initial ne reconnaît que les 26 majuscules et tentons de lui adjoindre 10 nouvelles classes correspondant aux chiffres.

De par sa structure modulaire, le classifieur à base de prototypes peut aisément être adapté afin de traiter de nouvelles classes. Contrairement aux classifieurs discriminants qui nécessitent un ré-apprentissage complet, les classifieurs générateurs sont incrémentaux par nature ; le modèle de chaque classe étant estimé sur des données appartenant à cette seule classe. De plus, une fois estimés, les nouveaux modèles sont directement inclus dans le classifieur et la décision est prise en considérant l'ensemble des modèles.

L'adaptation du classifieur hybride est ensuite effectuée. Dans un premier temps, la matrice de confusion étendue est déterminée sur l'ensemble d'estimation en considérant toutes les classes (ici les 26 majuscules initiales et les 10 « nouvelles » classes de chiffres soient 36 classes). Puis, les confusions croisées entre classes initiales et nouvelles classes sont détectées, comme précédemment. Enfin, les PNN correspondant aux confusions croisées sont entraînés et éventuellement activés.

	Classifieur générateur	**Classifieur hybride**	
Taux de reco	90.3 %	**M,M C,C**	**M,M C,C M,C**
		91.5 %	**95.4 %**
RND	-	**63**	**79**

Tableau 3 : Taux de reconnaissance du classifieur hybride incrémental (base de test : $B_{VC} \cup B_{TE}$).

Le tableau 7 précise les taux de reconnaissance du classifieur hybride dans les cas suivants :

- Classifieur générateur seul : le taux de reconnaissance chute drastiquement en raison des nombreuses confusions croisées entre lettres majuscules et chiffres comme (O,0), (I,1), (S,5) et (Z,2) par exemple.
- Classifieur hybride initial : les confusions entre majuscules (M,M) d'une part et chiffres (C,C) d'autre part sont traitées par l'étage discriminant mais les confusions croisées sont ignorées.
- Classifieur hybride adapté : toutes les confusions sont prises en compte, confusions croisées (M,C) inclues. Le résultat est très concluant : le taux d'erreur a été divisé par deux (comparé à celui du classifieur générateur).

3.1.5 Conclusions

Encadrements : 1 stage (Dewaele G.), 1 projet (Da Silva I.)

Nous avons présentés dans ce chapitre l'ensemble de nos travaux en reconnaissance de caractères isolés. Ces recherches ont été conduites entre 1995 (début du Doctorat) et 2002.

L'algorithme SMAC, en se libérant des habituels problèmes d'initialisation rencontrés par les techniques de clustering, a révélé ses capacités à produire une excellente partition des références. Son originalité repose sur trois points en particulier. L'adjonction d'une phase de modification des prototypes, à l'origine de l'algorithme SMAC2, a permis d'améliorer la qualité de la modélisation, augmentant d'autant les performances de classification. Le système est par essence très modulaire.

Nous avons ensuite combiné deux classifieurs, l'un dédié aux données dynamiques et à l'analyse de l'information gestuelle ; l'autre, aux données statiques et au traitement de l'information graphique. Leur complémentarité ont permis d'élaborer des algorithmes de fusion particulièrement efficaces. Les résultats obtenus étaient au meilleur niveau international comme l'a montré l'analyse comparative présentée dans [R03[48]].

[48] Ratzlaff E.H., Reports and survey for the comparison of diverse isolated character recognition results on the UNIPEN database, International Conference on Document Analysis and Recognition, (1), pp 623-628, 2003.

Par la suite, au lieu de changer la représentation de données, nous avons modifié le type de classifieurs. En combinant un classifieur générateur avec des classifieurs discriminants locaux, nous réussissons à raffiner les frontières entre classes confusives. Cette méthode présente l'avantage d'être incrémentale en architecture et en données. De plus, elle n'est plus guidée par les données mais générique et transposable à d'autres problèmes de classification. Plusieurs équipes de recherches ont d'ailleurs utilisé ce paradigme en reconnaissance d'écriture [VSV03[49], MSC05[50]], mais aussi, en reconnaissance d'objets [GRB07[51]]. Nous en verrons une autre application par la suite (§4.1).

Les classifieurs précédemment décrits, couplés à un algorithme de segmentation du flux de données, constituaient le noyau d'un système de reconnaissance de l'écriture scripte qui sera analysé en détail dans le prochain chapitre (§3.2.2). L'ensemble a aussi été intégré dans un système complet d'analyse et de reconnaissance d'expressions mathématiques manuscrites. Une application à vocation pédagogique (apprentissage du calcul assisté par ordinateur) a été conçue.

[49] Vuurpijl L., Schomaker L. & Van Erp M. Architecture for detecting and solving conflicts: two stage classification and support vector classifiers, International Journal on Document Analysis and Recognition, 5: 213-223, 2003.

[50] Milgram J., Sabourin R. & Cheriet M. Combining Model-based and Discriminative Approaches in a Modular Two-stage Classification System: Application to Isolated Handwritten Digit Recognition, Electronic Letters on Computer Vision and Image Analysis, 5(2):1-15, 2005.

[51] Grabner H., Roth P.M. & Bischof H., Eigenboosting: combining discriminative and generative information, IEEE Conference on Computer Vision and Pattern Recognition, à paraître, 2007.

3.2 RECONNAISSANCE DE TEXTES MANUSCRITS

3.2.1 Contexte

La reconnaissance automatique de l'écriture est aujourd'hui plus que jamais, un enjeu important. Il suffit d'observer la multiplication des appareils électroniques qui utilisent un stylo comme moyen d'interaction avec la machine pour s'en convaincre. Les ordinateurs, les assistants personnels, les livres électroniques et même les téléphones portables intègrent tous aujourd'hui un stylo. Leur succès est grandissant même s'ils ne disposent pas toujours d'un système de reconnaissance de l'écriture efficace. Les champs d'application de la reconnaissance de l'écriture ne sont toutefois pas cantonnés aux seuls objets nomades dépourvus de clavier. La compréhension de l'écriture permet l'automatisation de tâches laborieuses comme la lecture des montants de chèques, des adresses postales ou des formulaires. Il existe d'ores et déjà pour ces applications, des systèmes grand public réellement fiables.

En 2000, quand nous avons initié ces travaux, les systèmes de reconnaissance d'écriture commercialisés (*Calligrapher*, *Graffiti*. . .) étaient très contraignants. N'étant capables de traiter qu'un style d'écriture fixe (Figure 5.a), ils obligeaient l'utilisateur à ré-apprendre à écrire. A l'opposé, les systèmes développés en laboratoire étaient susceptibles de reconnaître l'´ecriture naturelle (cursive Figure 5.c), mais leurs performances n'étaient pas suffisantes pour envisager une commercialisation. En préférant reconnaître l'écriture scripte (Figure 5.b), nous avions sélectionner un niveau de contrainte intermédiaire puisque nous libérions l'utilisateur de la contrainte de ré-apprentissage, en lui imposant seulement d'écrire en levant le stylo entre chaque caractère. Ce choix nous permettait de profiter pleinement de notre expertise en reconnaissance de caractères isolés (§2.1). toutefois, en raison de la très grande variabilité du signal écrit, nous pouvions craindre que les performances d'un système apte à reconnaître n'importe quelle écriture (omni-scripteur) soient décevantes. C'est pourquoi nous avons envisagé de spécialiser le système de lecture à "son" utilisateur (mono-scripteur).

(a) (b) (c)

Figure 5. Catégories d'écriture : contrainte (a), scripte (b) et naturelle (cursive) [TSW90[52]].

Pour expliquer le déroulement du processus de lecture, les chercheurs en psychologie cognitive apparentent le cerveau du lecteur à un système informatique de traitement de l'information qui comprend dans sa version minimale un processeur, plusieurs mémoires, des organes récepteurs et d'éventuels effecteurs [BC95[53]]. Les modèles cognitifs de lecture développés par les spécialistes ne cherchent pas seulement à modéliser ce traitement de l'information, mais tentent également de modéliser l'importance du contexte et des effets contextuels observés chez l'Homme (effet de supériorité du mot, effet d'amorçage… [T91[54]]) dans la reconnaissance visuelle des mots. En effet, la pertinence d'un modèle cognitif de lecture est précisément évaluée en fonction de son aptitude à rendre compte des différents

[52] Tappert C., Suen C. Y. & Wakahara T., The state of the art in on-line handwriting recognition, IEEE Transaction on Pattern Analysis and Machine Intelligence, 12(8):787–808, 1990.
[53] Baccino T. & Cole P., La lecture experte, Presse universitaire de France, 1995.
[54] Taft M., Reading and mental lexicon, Erlbaum Edition, 1991.

effets contextuels. Le modèle le plus complet, appelé modèle à triple voie (sémantique, phonologiques [C78[55]]), n'a pas encore été exploité. Quelques systèmes [C97[56], P00[57]] exploitent le modèle d'activation interactive [MR81[58]]. La plupart des systèmes ([PS00[59], CA04[60], C00[61]]…), dont le notre, se basent sur le modèle d'activation vérification proposé par [PNM82[62]]. Dans ce modèle, le stimulus visuel d'entrée déclenche l'activation de certains mots du lexique (silhouette identique, orthographe proche…). Les mots ainsi activés constituent une liste de mots candidats. À l'étape de vérification, cette liste de mots est confrontée aux informations données en entrée (le stimulus de départ) afin de déterminer le meilleur candidat. La probabilité d'une réponse est la somme pondérée de la probabilité de réponse correcte basée sur une évidence lexicale et celle basée sur une évidence alphabétique.

La taxinomie des systèmes de reconnaissance de mots sépare les approches globales des approches analytiques. Les premières, qui exploitent l'effet de supériorité du mot, considèrent ce dernier dans son ensemble et cherchent à déterminer les caractéristiques [PSW97[63]] qui permettent de le discriminer des autres mots. Elles ne fonctionnent que sur des lexiques de petite taille. Les approches analytiques reposent sur une segmentation du mot à reconnaître. L'entité de base étant alors le graphème. L'intérêt d'un tel découpage est de ne concevoir que des modèles de graphèmes ou de lettres. Ces modèles sont en nombre limité (26 modèles pour les minuscules par exemple) et permettent de modéliser tous les mots quelque soit la taille du lexique utilisé. Les méthodes qui prédominent pour la classification des graphèmes ont déjà été présentées précédemment (§3.1.1).

3.2.2 Reconnaissance de textes

Encadrements : 1 thèse (Oudot L.), 1 stage (Oudot L.)

3.2.2.1 Système de lecture

La structure générale du moteur de lecture (Figure 6) s'inspire du modèle d'activation vérification. Le système se présente sous la forme d'une série d'experts d'encodage qui permettent d'extraire des informations géométrique et morphologique du texte écrit par l'utilisateur. Ces experts fournissent des informations probabilistes au niveau *stroke* (défini portion de tracé comprise entre deux levers de stylo). Un premier expert de pré-traitement, non représenté sur la figure, segmente le flux d'entrée en lignes car notre système traite les

[55] Coltheart M., Lexical access in simple reading task, Underwood, Strategies of information processing, Academic Press, 1978.
[56] Coté M., Utilisation d'un modèle d'accès lexical et de concepts perceptifs pour la reconnaissance d'images de mots cursifs, Thèse de doctorat, ENST, 1997.
[57] Pasquer L., Conception d'un modèle d'interprétation multicontextuelle, application à la reconnaissance en-ligne d'écriture manuscrite », Thèse de Doctorat, Université Rennes I, 2000.
[58] Mac Lelland J.L. & Rumelhart D.E., An interactive activation model of context effects in letter perception, Psychological Review, Vol. 88, pp 375–407, 1981.
[59] Plamondon R. & Srihari S.N. On-line and off-line handwriting recognition: a comprehensive survey, IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(1):63-84, 2000.
[60] Carbonnel S. & Anquetil E., Modélisation et intégration de connaissance lexicales pour le post-traitement de l'écriture manuscrite en-ligne, Congrès Reconnaissance des Formes et Intelligence Artificielle, Vol. 3, pp 1313–1322, 2004.
[61] Connel S.D., On-line Handwriting Recognition Using Multiple Pattern Class Models, PhD thesis, Department of Computer science and engineering, Michigan State University, 2000.
[62] Paap K., Newsome S.L., Mac Donald J.E., Schvaneveldt R.W., An activation-verification model for letter and word recognition: The word superiority effect, Psychological Review, Vol. 89, pp 573–594, 1982.
[63] Powalka R.K., Sherkat N. & Whitrow R.J., Word shape analysis for a hybrid recognition system, Pattern Recognition, 30(3): 421–445, 1997.

informations ligne de texte par ligne de texte. Toutes ces informations, aussi bien au niveau local qu'au niveau global, permettent d'activer une liste de mots du lexique. La cohérence de chaque mot hypothèse de la liste est ensuite évaluée par un moteur probabiliste qui valide la meilleure hypothèse pour la retranscription. Nous avons évalué ce système à l'aide de deux lexiques français de taille différente. Le premier $Dico_{fr}$ contient 185000 mots et couvre un très large champ d'applications. Le second $Dico_{rd}$ contient les 8000 mots les plus fréquents de la langue française.

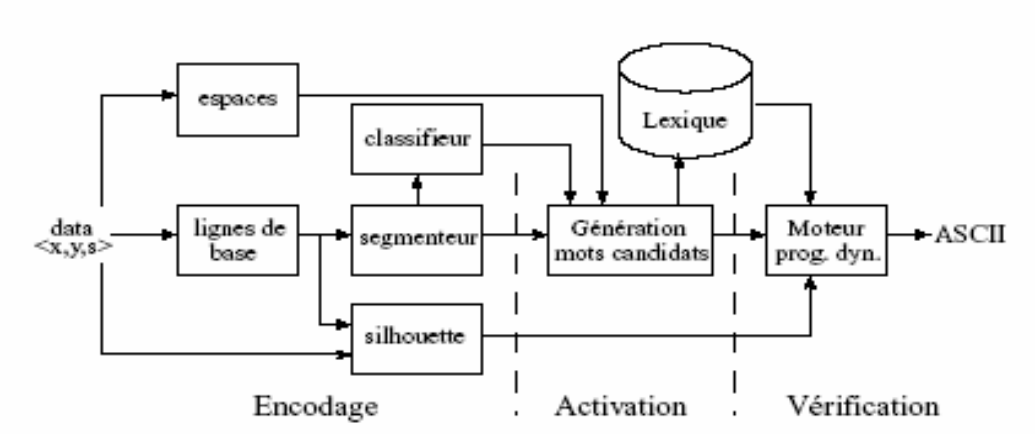


Figure 6. Structure générale du système de lecture.

Pour les expérimentations, nous avons collecté une base de 90 textes écrits par 54 scripteurs différents réunissant 26000 lettres, 5400 mots et 3000 signes de ponctuation. La Figure 7 présente deux exemples de texte. Cette base a été divisée en deux parties égales : une première base B_A pour l'estimation des paramètres du système de lecture et une seconde, B_T, pour son évaluation. Chaque base contient 45 textes. La base d'apprentissage comprend 25 scripteurs dont 14 communs avec la base de test. La base de test quant à elle, contient 29 scripteurs et inclut donc 15 scripteurs totalement inconnus de la base d'apprentissage. Le fait de ne pas totalement différencier les scripteurs entre les deux bases permet de mieux évaluer les résultats et la pertinence du système et de comparer simultanément son comportement dans les cadres omni-scripteur et multi-scripteur. Les taux de bonne classification donnés par la suite ont tous été évalués sur la base de test B_T.

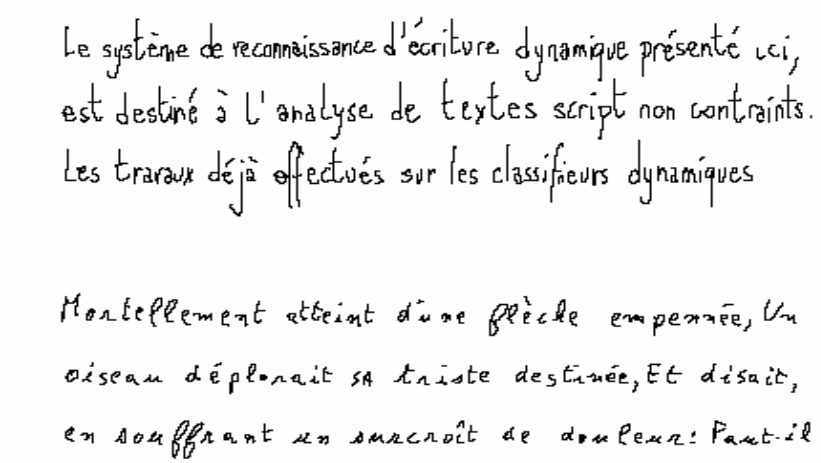


Figure 7. Exemples de texte (respectivement reconnus à 99% et 70% par le système de lecture).

3.2.2.2 Experts d'encodage

Détection des lignes de base
Cet expert recherche les sauts de lignes pour segmenter le flux de données d'entrée en ligne. Les techniques les plus courantes utilisent le contour du texte [SR98[64]] ou ajustent des lignes paraboliques parallèles par un processus itératif. Notre expert approxime les lignes d'assises par morceau avec des lignes droites. Pour cela, il recherche les strokes représentant des lettres médianes (*a, c, e, i, n ...*) pour les relier ensuite par des lignes droites. On relie le bas des boîtes englobantes de ces derniers pour obtenir la ligne de base et le haut des boîtes pour obtenir la ligne de corps (Figure 8).

Caractérisation de la silhouette
Le rôle de cet expert est de déterminer la silhouette des mots en utilisant les lignes de bases et d'extraire ainsi des informations globales [PSW97]. Cet expert estime pour chaque stroke s_i deux probabilités correspondant à la présence d'une hampe $p(h|s_i)$ et d'un jambage $p(j|s_i)$. De plus, à partir de ces deux informations, on attribue au stroke une classe d'activation Act_i parmi les quatre suivantes : médiane (*m*), hampe (*h*), jambage (*j*) et hampe/jambage (*f*). L'estimation de ces probabilités est basée sur l'utilisation des lignes d'assises calculées par l'expert précédent. La classe d'activation Act_i est déterminée par un réseau de neurones. Le taux de bonne classification dépasse 94%.

Détection des espaces
Cet expert fournit, pour chaque arc inter-strokes a_i, une probabilité $p(e_j|a_i)$ d'être un espace séparant deux mots distincts. Cette information est précieuse pour la localisation des mots dans une ligne de texte. Un réseau de neurones minimise la probabilité d'erreur de classification à partir de la distance horizontale inter-strokes. Afin de disposer d'une probabilité en sortie de ce réseau, la cellule de sortie utilise une fonction de transfert linéaire saturée entre 0 et 1. Le taux de bonnes classifications obtenu est de 95,4%. On constate que le détecteur a tendance à sous segmenter les mots en fusionnant deux ou plusieurs mots consécutifs. Mais ce phénomène est avant tout caractéristique de l'écriture scripte ; nombre d'utilisateurs séparant autant les lettres d'un même mot que les mots entre eux (égalité des espaces intra-mot et inter-mots).

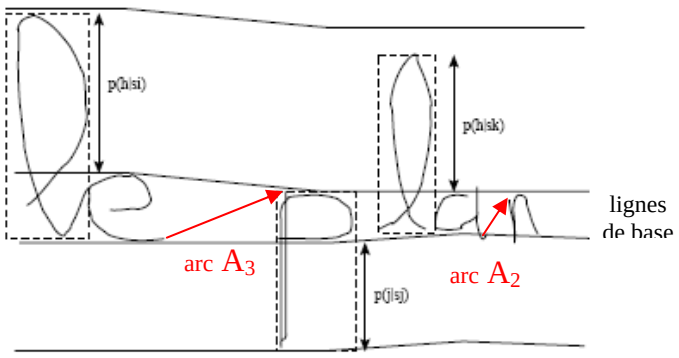


Figure 8. Encodage du signal écrit.

[64] Senior A. W. & Robinson A. J., An off-line cursive handwriting recognition system, IEEE Transaction on Pattern Analysis and Machine Intelligence, 20(3) :309–321, 1998.

Segmenteur

Les lever de stylo, tout comme les strokes, contiennent une grande quantité d'informations. Comme [GMC97[65]], on utilise un réseau de neurones pour classifier ces arcs inter-strokes. A partir des données d'entrée, on construit un graphe non orienté $G=(S,A)$ avec un ensemble de sommets constitués des strokes $S=\{s_1, s_2, \ldots s_n\}$ et des arêtes $A=\{(s_1, s_2), (s_2, s_3) \ldots (s_{n-1},s_n)\}$ correspondant aux arcs reliant deux strokes consécutifs du point de vue temporel. On distingue cinq classes d'arcs : intra-lettre (A_1), inter-lettres & intra-mot (A_2), inter-mots (A_3), diacritique (A_4) et ponctuation (A_5). 32 paramètres servent à caractériser les arcs inter-strokes [O03[66]]. Ils sont estimés en considérant pour deux strokes consécutifs s_{n-1} et s_n, leurs coordonnées, leurs boîtes englobantes, leurs positions par rapport aux lignes de bases et le lever de stylo (défini comme le segment joignant le dernier point de s_{n-1} au premier point de s_n). L'idéal étant de trouver des paramètres invariants aux scripteurs (paramètres omni-scripteurs). L'analyse factorielle de cet ensemble de paramètres n'ayant pas produit de frontières aisément modélisables, nous nous sommes tournés vers les méthodes de sélection de paramètres de type *wrapper* (contrôlées par un classifieur). La stratégie utilisée est dite FBSS (*Forward Backward Sequential Selection*) et se rapproche des méthodes de *Floating Search* car, partant d'un sous-ensemble de départ d'une dizaine de paramètres tirés aléatoirement, elle alterne des phases *Forward* (ajout de paramètres augmentant le taux de reconnaissance) et *Backward* (suppression) jusqu'à stabilisation du taux de reconnaissance. Nous avons comparé plusieurs classifieurs : direct, bayésien et neuronal. Le meilleur classifieur obtenu est un réseau MLP (32-20-5). Le taux de reconnaissance est de 92,3% pour la première réponse (top_1) et de 99,1% pour le top_2.

Classifieur de caractères

Le classifieur de caractères évalue pour chaque stroke ou combinaison de strokes les probabilités d'appartenance aux 62 classes (26 minuscules, 26 majuscules et 10 chiffres) : $p(c|data)$, $c \in \Gamma$, $\Gamma=\{A,\ldots,Z,a,\ldots,z,0\ldots,9\}$. Le classifieur à base de prototypes est de type 1-ppv. L'exemple inconnu sera comparé au corpus de prototypes de la classe correspondante. La métrique utilisée est la distance élastique. Les probabilités sont obtenues par l'utilisation de la règle du *softmax*. Les bases de prototypes du classifieur sont obtenues par l'algorithme de clustering SMAC (§3.1.2). Une première base de prototypes extraite de la base UNIPEN a donné des résultats décevants. Pour améliorer ces résultats, deux autres bases ont été crées à partir de la base de texte d'apprentissage B_A. La première, extraite des textes écrits par les scripteurs qui n'apparaissent pas dans la base de test, permet d'évaluer le comportement du classifieur en contexte omni-scripteur. La seconde, extraite de l'ensemble des textes de la base d'apprentissage, correspond au contexte multi-scripteur. On constate (Tableau 4) les améliorations de performance successives.

Base de prototypes	UNIPEN	Omni-scripteur	Multi-scripteur
Taux de reconnaissance	78.9%	84.8%	88.7%

Tableau 4. Performances du classifieur dynamique sur la base B_T en fonction de la base de prototypes.

[65] Gader P.D., Mohamed M. & Chiang J.-H., Handwritten Word Recognition with Character and Inter-Character Neural Networks, IEEE Transaction on systems, Man and Cybernetics, 27(1):158–164, 1997.

[66] Oudot L., Fusion d'informations et adaptation pour la reconnaissance de textes manuscrits dynamiques, Thèse de doctorat de l'Université Pierre et Marie Curie-Paris 6, 2003.

3.2.2.3 Moteur de lecture

Génération du treillis de segmentation
Une technique classique pour retrouver la segmentation d'un texte, est de générer un treillis d'hypothèses de segmentation [PSW97, HS01[67]] et de laisser aux modules suivants (en général le classifieur de caractère et/ou le lexique) la prise de décision. L'inconvénient avec cette technique est de trouver un bon compromis entre le nombre d'hypothèses générées et la vitesse de traitement. Générer beaucoup d'hypothèses permet d'augmenter la probabilité d'obtenir la bonne segmentation mais risque de perturber et ralentir le système par un trop grand nombre d'hypothèses. Il faut donc réussir à générer un minimum d'hypothèses les plus pertinentes possibles.

Pour une ligne de texte « moyenne » comportant 40 strokes, le treillis d'hypothèse généré pour être certain d'avoir la bonne segmentation sera de taille $5^{40} \cong 10^{27}$. Ne conserver que les deux meilleurs résultats (top_2) du segmenteur permet, on l'a vu, d'atteindre 99,1% d'apparition de la bonne segmentation[68]. Mais un tel treillis possède encore environ $2^{40} \cong 10^{12}$ hypothèses pour une ligne moyenne et est donc inutilisable. Au contraire, en ne considérant que la première réponse, on génère une hypothèse unique, mais le taux d'apparition n'est que de 92,3%.

Pour réduire le nombre d'hypothèses de segmentation, nous utilisons le détecteur d'espaces inter-mots (classe A_3) pour fixer des points d'ancrage dans la ligne. Ces derniers découpent le treillis de segmentation en plusieurs sous-treillis plus petits et cassent la combinatoire. En considérant les arcs ayant une probabilité p_{espace}=1 comme étant des point d'ancrage, on retrouve 50% des espaces inter-mots avec moins de 1% de fausses détections. Le treillis de segmentation est ensuite généré progressivement en ajoutant les types d'arcs les plus probables jusqu'à atteindre N_{hypo} hypothèses de segmentation (Figure 9).

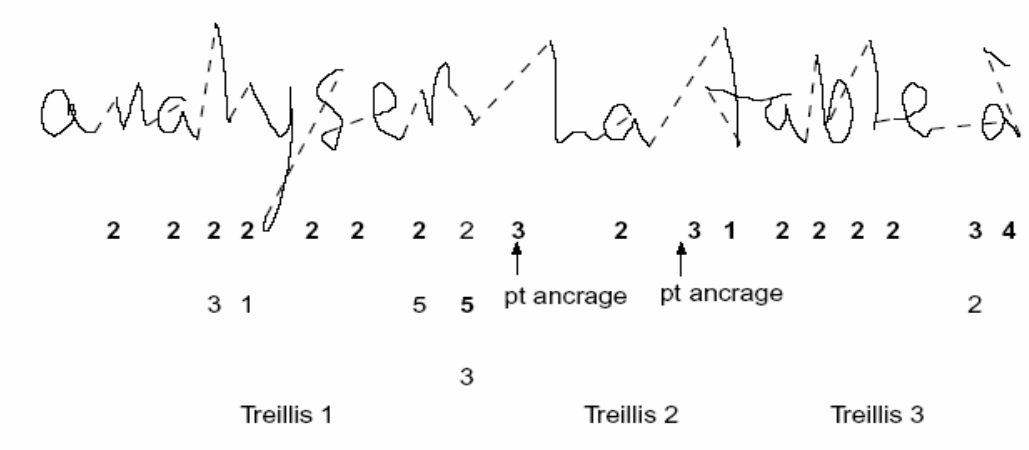


Figure 9. Exemple de treillis de segmentation avec points d'ancrages (26 hypothèses).

La pertinence d'une classe d'arc pour un arc donné est calculée en fusionnant les probabilités *a priori* et *a posteriori* du segmenteur :

$$p(A_i|data) = \sum_{a=A_1}^{A_5} p(A_i|a)\, p(a|data)$$

[67] Henning A. & Sherkat N., Cursive script recognition using wildcards and multiple experts, Pattern Analysis & Applications, 4 :51–60, 2001.
[68] Remarque : le taux d'apparition correspond à l'existence de la segmentation réelle dans le treillis et non au taux de bonne segmentation.

Avec *p(a|data)* la probabilité *a posteriori* d'une segmentation et $p(A_i|a)$ la probabilité *a priori* évaluée à partir de la matrice de confusion du segmenteur. Cette dernière est estimée sur la base d'apprentissage B_A et synthétise le comportement du segmenteur observé sur cette base. Si la base d'apprentissage est statistiquement représentative, il y a de fortes probabilités que ce comportement soit généralisable. Cette méthode donne de très bons résultats : avec un treillis limité à N_{hypo} =500 hypothèses, on obtient 98,4% de bonne apparition de la segmentation et pour N_{hypo} =2000, plus de 98,8%.

Activation des mots candidats

L'activation des mots candidats du lexique (modèle d'activation vérification) est introduit par une recherche lexicale dans une sous partie restreinte du lexique principal. En effet, la recherche ne s'effectue que sur les mots de même longueur, l'organisation du lexique étant basée sur le nombre de caractères des mots.

Les mots hypothèses sont obtenus en conservant pour chaque lettre hypothèse, la classe la plus probable du classifieur de caractères et possédant la même silhouette que celle observée par le détecteur de silhouette. Une recherche lexicale en tolérant N_{err} caractères de différence constitue la liste des mots lexicalement corrects. Le seuil N_{err} = E(N_{car}/2) (où N_{car} est le nombre de caractères du mot hypothèse) a été choisi de manière à avoir le meilleur compromis entre le nombre de mots générés et l'apparition du bon mot dans cette liste. Nous obtenons 97,7% de bonne apparition du mot réel sur la base de test avec une moyenne de 50 mots candidats par proposition. Les 2,3% d'erreur restants sont dus à de trop nombreuses erreurs de classification, surtout sur les mots courts où plus de la moitié des lettres sont fausses.

Retranscription du texte

La retranscription d'une ligne de texte se décompose en deux étapes. La première assigne à chaque mot hypothèse une probabilité de cohérence et la seconde effectue une recherche de la ligne la plus probable.

La probabilité d'un mot de la liste *p(mot|data)* est le produit de la probabilité des caractères le composant *p(car|data)* multiplié par la probabilité de sa segmentation *p(seg|data)*. La probabilité de tous les caractères *p(car|data)* composant un mot revient à estimer les probabilités de chaque caractère *p(car(i)|data)* connaissant les données d'entrée :

$$p(car|data) = \prod_{i=1}^{N_{car}} p(car(i)|data)$$

L'estimation de la probabilité des caractères tient compte des probabilités *a posteriori* du classifieur de caractères ainsi que du détecteur de silhouette et des probabilités *a priori* déduites de la matrice de confusion de ces deux experts.

$$p(car(i)|data) = \sum_{c\in\Gamma, s\in\{m,h,j,f\}} p(car(i)|(c,s))\, p((c,s)|data)$$

Le terme *p((c,s)|data)* correspond à la probabilité d'observer le caractère *c* ayant la silhouette *s* connaissant les données d'entrée. Ces informations sont données par deux experts : classifieur de caractères et caractérisation de la silhouette. En considérant les deux sources indépendantes, il vient : $p((c,s)|data) = p(c|data)\, p(s|data)$. Le terme *p(car(i)|(c,s))* est déduit de la matrice de confusion.

La probabilité de la segmentation est déduite des probabilités calculées précédemment lors de la génération du treillis (fusion des probabilités *a priori* et *a posteriori* du segmenteur) :

$$p(seg|data) = \prod_{i=1}^{N_{car}} p(s|a_i)$$

Chaque mot de chaque hypothèse de segmentation d'une ligne a donc une probabilité associée et une position de début et de fin (en numéro de stroke) dans la ligne. Ceci permet de générer un treillis de mots hypothèses. En utilisant un algorithme de programmation dynamique on retrouve le chemin le plus probable à l'intérieur du treillis de mots hypothèses. La force de cet algorithme réside dans le caractère global de la recherche. Le chemin le plus probable retrouvé est une combinaison de plusieurs hypothèses de segmentation (Figure 10).

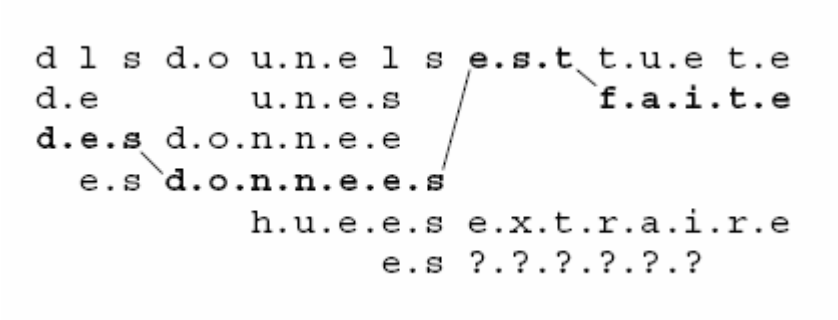


Figure 10. Exemple de treillis de mots
(Les points correspondent à des espaces intra-mots, les blancs à des espaces inter-mots).

3.2.2.4 Résultats expérimentaux

Le système complet de lecture automatique de texte décrit précédemment est aisément portable puisque son encombrement mémoire ne dépasse pas 500Ko (en incluant les codes, le lexique et la base de prototypes). Il traite en moyenne 6 mots par seconde (sur un Pentium IV cadencé à 1.8 GHz) Nous avons effectué deux expériences. Dans la première, nous considérons connue la segmentation du texte en mots. Il s'agit donc d'une évaluation des performances du moteur en reconnaissance de mots isolés. Comme on pouvait s'y attendre (Tableau 5), les taux de reconnaissance sont élevés (environ 90%) et augmentent lorsque la taille du lexique diminue et que la base de prototypes se « rapproche » des utilisateurs (contexte multi-scripteur). Dans la seconde expérience, nous laissons le moteur de lecture segmenter le texte. Les taux de reconnaissance en mots et en lettres obtenus pour N_{hypo}=500 sont plus faibles que précédemment. On retrouve par contre les deux phénomènes déjà constatés. L'utilisation d'un très grand lexique ne fait chuter le taux de reconnaissance en mots d'environ 3%. Ce dernier s'améliore en contexte multi-scripteur, ce qui montre l'intérêt de l'adaptation du moteur de lecture à son utilisateur que nous présentons dans la suite.

Base de prototypes	Omni-scripteur		Multi-scripteur	
Segmentation	connue	inconnue	connue	inconnue
Dico fr	89%	72%	91%	75%
Dico rd	91%	75%	94%	78%

Tableau 5. Taux de reconnaissance (mots) du moteur de lecture.

3.2.3 Adaptation au scripteur[Annexe 3]

Encadrements : 1 thèse (Oudot L.), 2 stages (Moises A., Pajadon A.)

3.2.3.1 Principe

L'adaptation du lecteur à l'écriture du scripteur a été mise en lumière par les travaux en psychologie perceptive. On peut remarquer que dans le cas d'une écriture difficile à lire, on lira plus facilement un mot écrit par un individu si l'on a auparavant lu un autre mot écrit par le même individu. Il s'agit d'un effet d'amorçage graphémique. Il est facile de constater ce phénomène dans la pratique lors de la lecture d'un texte très peu lisible. Quand un mot est illisible ou ambiguë, nous recherchons instinctivement dans ce texte, d'autres mots que nous pouvons lire pour comprendre ce premier mot. Nous apprenons donc les caractéristiques du scripteur à partir de mots que nous avons pu lire, pour ensuite utiliser ces nouvelles connaissances sur le reste des mots auparavant illisibles.

Nous avons pu constater lors de l'étude précédente, la base de prototypes du classifieur, aussi riche soit-elle, n'est pas encore assez représentative. D'autre part, l'utilisation des caractères tirés de la base d'apprentissage pour une utilisation en mode multi-scripteur, permet d'améliorer les taux de reconnaissance. Nous faisons donc l'hypothèse qu'une spécialisation du système à un utilisateur donné (mode mono-scripteur), aurait des conséquences encore plus bénéfiques. Concrètement, il existe deux sources d'erreurs :

- Le graphème est absent de la base omni-scripteur, il faut donc enrichir cette dernière : le graphème doit être stocké dans la base de prototypes (Ajout).
- Le graphème est confusif : pour un scripteur donné, les modèles de la base omni-scripteur doivent être supprimés afin d'éviter toute confusion (Suppression).

Dans la littérature, on considère deux stratégies d'adaptation. Les systèmes hors-ligne ou *batch* [CJ02[69], L02[70]] sont adaptés une fois pour toute avant d'être utilisés. Ils sont spécialisés, à l'aide d'un apprentissage supervisé (*enrolement*) nécessitant un ensemble de données étiquetées fourni par l'utilisateur. Au contraire, les systèmes à adaptation continue (en-ligne) [GHA92[71], PM97[72], VLK02[73]] évoluent en permanence pendant la phase d'utilisation, sous la supervision de l'utilisateur. Nos proposons ici de réaliser cette adaptation de façon non supervisée, autrement dit, sans solliciter l'utilisateur. Un problème se pose toutefois quelque soit la stratégie d'adaptation : celui de la capacité de généralisation du modèle. En effet la multiplication des paramètres des modèles neuronaux, markoviens, flous ou hybrides rend la convergence de leur apprentissage quasi impossible lorsque le nombre d'exemples disponibles est limité. La solution la plus simple reste le classifieur *k*-ppv à prototype qui peut facilement modéliser une très grande variabilité de l'´ecriture quelque soit le nombre d'exemples disponible. Un unique exemple suffit même à adapter le système.

[69] Connell S. & Jain A.K. Writer adaptation of on-line handwriting models, IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(2): 329-342, 2002.

[70] Li H., Traitement de la variabilité et développement de systèmes robustes pour la reconnaissance de l'écriture manuscrite en-ligne , Thèse de doctorat de l'Université Pierre et Marie Curie-Paris 6, 2002.

[71] Guyon I., Henderson D., Albrecht P., Le Cun Y.& Denker J., Writer independent and writer adaptative neural network for on-line character recognition, From Pixel to Features III, 1992.

[72] Platt J.C. & Matt N.P., A constructive RBF network for writer adaptation, Advances in neural information processing systems, Vol. 1, pp. 765–771, 1997.

[73] Vuori V., Laaksonen J. & Kangas,J. Influence of erroneous learning samples on adaptation in on-line handwriting recognition, Pattern Recognition, 35(4): 915-925, 2002.

3.2.3.2 Adaptation supervisée

Pour évaluer la pertinence de l'adaptation au scripteur, nous avons utilisé les étiquettes des textes pour effectuer une adaptation supervisée. Les formes du texte sont classifiées l'une après l'autre. Les hypothèses de classification (qui correspondent aux labels *top1* du classifieur de caractères) sont comparées aux étiquettes. Si elles ne correspondent pas, la forme en question est ajoutée à la base utilisateur (Figure 11). Cet ajout est réalisé ligne par ligne. Dans une première expérience, nous supposons la segmentation des textes en mots et lettres connue. L'adaptation supervisée décrite ci-dessus fait décroître très rapidement le taux d'erreur moyen qui atteint 1% après 50 mots écrits et est inférieur à 0.5% pour 150 mots. A l'issue de la phase d'adaptation, la taille de la base de prototypes n'a crue, en moyenne, que de 4%. Le taux de reconnaissance du système de lecture réel (segmentation inconnue) avec une adaptation par *enrolement* (utilisation de la base de prototype utilisateur déterminé auparavant) atteint 12% d'erreur au lieu des 28% du système seul sans adaptation.

3.2.3.3 Adaptation non-supervisée

L'adaptation non-supervisée est définie par le fait que nous ne possédons plus l'étiquette du texte. Il s'agit d'une situation réelle d'utilisation du système de lecture qui s'adapte au fur et à mesure à l'utilisateur sans le solliciter. Cette fois, les informations délivrées par le classifieur de caractères et le correcteur lexical sont comparées pour trouver les prototypes à ajouter.

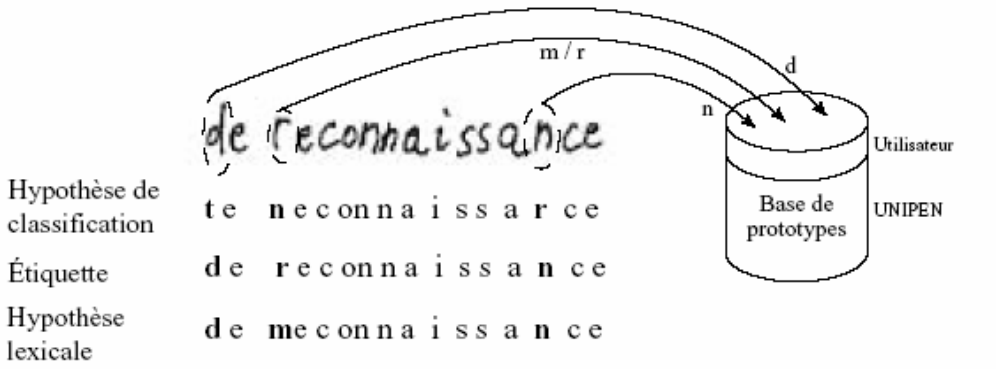


Figure 11. Ajout de prototypes dans la base utilisateur (adaptation supervisée et non supervisée).

Ajout systématique

Dans la stratégie d'adaptation par ajout systématique, on considère que le taux d'erreur du correcteur lexical est nul. Donc, à chaque erreur (différence entre la classification et l'hypothèse lexicale) le caractère est ajouté à la base de prototypes utilisateur (Figure 11). Les erreurs de correction lexicale cumulées à celles de la segmentation, conduisent à l'ajout de prototypes dans de mauvaises classes. Ces mauvais ajouts entraînent de nombreuses erreurs de classifications par la suite. Le gain dû à l'adaptation est faible (Tableau 6) : le taux de reconnaissance du système adapté dépasse les 75% après 50 mots écrits.

	Taux d'erreur			Base utilisateur
Mots	50	100	150	
Sans adaptation		28%		100%
Systématique: min	0%	1,9%	2%	+2%
moy	25%	23%	23%	+6%
max	53%	73%	51%	+14%
Conditionnel: min	0%	0%	2%	+1%
moy	**22%**	**20%**	**17%**	**+2%**
max	71%	58%	43%	+3%

Tableau 6. Adaptation non supervisée : évolution du taux d'erreur et de la base utilisateur.

Ajout conditionnel

La procédure d'ajout systématique étant peu performante, il semble nécessaire d'étudier le comportement de l'analyseur lexical pour n'ajouter que des prototypes utiles. Deux conclusions s'imposent lors de cette étude :

- La fiabilité de l'analyseur lexical augmente avec le nombre de lettres composant le mot.
- La fiabilité de l'analyseur lexical diminue lorsque le nombre d'erreur (lettres du mot mal classées) de classification augmente.

Le moteur de reconnaissance estime pour chaque mot une probabilité d'apparition d mot connaissant les données $p(mot|data)$ que nous appellerons de cohérence lexicale (PCL). L'ajout conditionnel consiste à n'utiliser, pour l'ajout de prototype, que les mots dont la PCL est supérieure à un seuil (déterminé par recherche exhaustive). Les meilleurs résultats sont présentés dans le Tableau 6. L'ajout conditionnel permet de réduire considérablement les mauvais ajouts de prototypes de l'ajout systématique. Ceci conduit à n'augmenter la taille de la base utilisateur que de 2%, tandis que le taux de reconnaissance dépasse les 80% après 100 mots écrits.

G*estion dynamique des prototypes*

Cette stratégie a deux objectifs : la suppression des prototypes erronés dus aux erreurs lors de l'ajout conditionnel (erreur du correcteur lexical) et la réduction de la taille de la base utilisateur en supprimant les prototypes inutiles [VLK02] afin d'accélérer la vitesse de classification. On attribue à chaque prototype (de la base omni-scripteur comme de la base utilisateur) une adéquation Q initialisée à la valeur $Q(0)=1000$. Cette adéquation sera modifiée à chaque occurrence du prototype suivant l'utilité de ce dernier dans le processus de classification en comparant l'hypothèse de classification à l'hypothèse lexicale. Pour le prototype i : $Q_i(n+1)= Q_i(n)+[C(t)-I(t)-N(t)]/F$ où C, I et N sont trois paramètres :

- C : Récompense (+) du prototype i quand l'hypothèse de classification est *C*orrecte.
- I : Pénalité (-) du prototype i quand l'hypothèse de classification est *I*ncorrecte.
- N : Pénalité (-) pour tous les prototypes *N*on utilisés de la classe.

F est la fréquence de la lettre correspondante dans la langue française. Les trois paramètres sont mutuellement exclusifs : à chaque occurrence du prototype, un seul paramètre est activé. Lorsque $Q_i = 0$, le prototype est éliminé. Après recherche exhaustive des paramètres *(C, I, N)* optimaux pour une combinaison avec la méthode d'ajout conditionnel, les meilleurs résultats sont obtenus avec le triplet (30, 200, 8). Comme on peut le voir (Tableau 7), cette méthode a permis de réduire la taille de la base de prototypes de 40% sans modification du taux de reconnaissance.

Nous allons maintenant étudier en détail l'évolution de l'adéquation de certains prototypes (Figure 7). Pour certains scripteurs, les prototypes omni-scripteur sont suffisants. Pour la classe 'a' (ligne 1, à gauche), 2 prototypes sont utilisés et donc l'adéquation des 45 autres décroît. Pour la classe 's' (ligne 1, à droite), 4 prototypes sont utiles (le scripteur à une écriture instable) et les 36 autres sont désactivés. Pour d'autres scripteurs (ligne 2), des prototypes utilisateurs (en gras) sont nécessaires. Au départ, un prototype omni-scripteur est utilisé et après quelques occurrences, un prototype utilisateur est ajouté (l'utilisateur s'est habitué à la tablette).

	Taux d'erreur	Base utilisateur
Sans adap.	28%	100%
Gestion dyn.	**17%**	**–40%**

Tableau 7. Gestion dynamique des prototypes : évolution du taux d'erreur et de la base utilisateur.

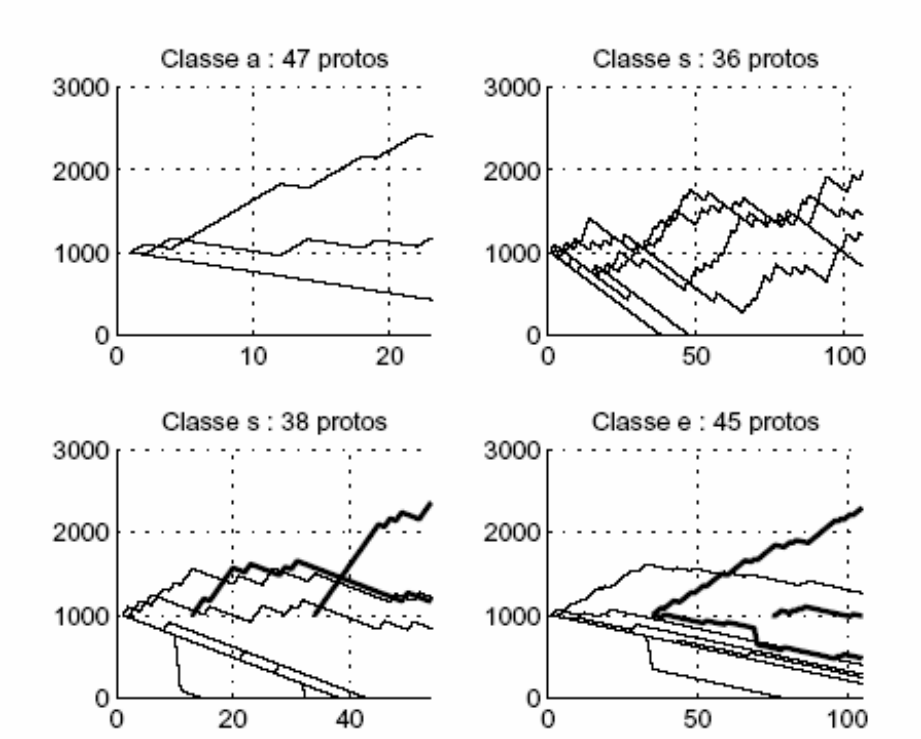


Figure 12. Evolution de l'adéquation des prototypes fonction du nombre d'occurrences.

3.2.3.4 Adaptation semi supervisée

Les performances du système de lecture réel (segmentation inconnue) utilisant la base de prototype modifiée par ajout conditionnel et gestion dynamique s'améliorent avec un taux d'erreur de 17% au lieu des 28% du système initial. Ainsi, l'adaptation non supervisée atteint presque les performances du système après enrolement (12%) sans aucune intervention de l'utilisateur. La dernière stratégie, appelée semi supervisée, combine l'adaptation supervisée du système sur quelques mots suivie par l'adaptation non-supervisée par gestion dynamique. Comme on peut le constater (Tableau 8), cette combinaison s'avère très efficace. Ainsi, imposer à l'utilisateur d'écrire une phrase de 30 mots permet d'améliorer les performances en portant le taux de reconnaissance à plus de 90%. Ces résultats pourraient encore s'améliorer en choisissant judicieusement ces mots, afin de s'assurer une bonne représentativité de l'ensemble des caractères à reconnaître (alors que dans la procédure actuelle, les 30 premiers mots écrits ont été utilisés pour l'adaptation supervisée).

Mot enrolement	Taux de reconnaissance	
	Après enrolement	100 mots
0	70 %	76 %
10	74 %	83 %
20	74 %	88 %
30	**74 %**	**90 %**
50	75 %	91 %

Tableau 8. Taux de reconnaissance, adaptation semi-supervisée

3.2.4 Conclusions

Encadrements : 1 stage (Michel-Sendis C.)

L'utilisation de concepts issus des recherches menées en psychologie perceptive tels que l'effet de la supériorité du mot, a permis de construire un système de lecture automatique de l'écrit efficace. L'encodage est réalisé par plusieurs experts simples fournissant des informations de type probabiliste à des niveaux d'abstraction différents. Le segmenteur utilise ces informations pour générer un treillis de segmentation dont la profondeur est limitée grâce à l'utilisation de points d'ancrage et d'informations *a priori*. Nous pouvons donc générer le treillis d'hypothèses le plus adapté à la puissance de calcul disponible. Le modèle d'accès lexical, direct, combine les informations géométriques (silhouettes des mots) et sub-lexicales (hypothèses de classification) pour générer une liste de mots candidats. Une programmation dynamique réalisée globalement sur l'ensemble des hypothèses permet de retranscrire le texte manuscrit. Les résultats obtenus dans le cadre omni-scripteur sont très encourageants : 75 % des mots sont reconnus avec un lexique de 8 000 formes. L'utilisation d'un très grand lexique de 180 000 mots ne fait chuter les performances que de 3 %.

Nous avons proposé plusieurs méthodes d'adaptation au scripteur. La première modifie la base de prototypes utilisateur par ajout conditionnel des allographes caractéristiques de son écriture. La seconde gère les prototypes dynamiquement, ce qui permet de réduire la taille de la base de prototypes pour ne garder que les plus pertinents. Cette diminution du nombre de paramètres entraîne une augmentation de la vitesse de traitement. Enfin, la stratégie semi-supervisée, en mettant à contribution l'utilisateur, permet d'atteindre très rapidement des performances élevées : 90% des mots sont bien reconnus.

Cette dernière méthode permet d'envisager l'extension de la reconnaissance à l'écriture naturelle, cursive. En effet, puisqu'il possible d'apprendre les modèles de lettres caractéristiques de l'écriture d'un utilisateur donné, pourquoi ne pas apprendre aussi les modèles de liaisons inter-lettres de cette même écriture ? En analysant plusieurs échantillons d'écriture cursive produits par un même scripteur, on constate qu'il existe des levers de stylos intra-mots relativement stables entre paires de lettres identiques. Ces levers pourraient constituer de nouveaux points d'ancrage de segmentation. Les portions d'écriture cursive comprises entre deux ancres seraient segmentées via des techniques classiques (recherche des minimas/maximas locaux du signal), ce qui permettrait d'identifier les modèles de lettres et de liaisons.

Nous avons choisi, par manque de temps, d'explorer une autre voie : le modèle de perception généralisé à triple voie de Coltheart. Nous espérions que la recherche de mots dans deux listes, lexicales et phonétiques améliorerait les performances du système de lecture. Malheureusement, les résultats obtenus n'ont pas été à la hauteur de nos espérances.

3.3 ANALYSE DE VISAGES

3.3.1 Contexte

L'analyse de visages est un domaine de recherche en perpétuelle expansion en raison de ses nombreuses applications : interaction homme-machine, indexation, biométrie, analyse comportementale … Elle inclut plusieurs thématique de recherche : la détection (y-a-t'il un visage dans l'image ?), la localisation (où se trouve le visage dans l'image ?), la reconnaissance (qui est la personne dont le visage a été détecté ?), l'authentification (la personne est-elle bien celle qu'elle prétend être ?) … Les difficultés auxquelles font face les systèmes d'analyse de visages sont multiples : forte variabilité des paramètres intrinsèques comme l'expression faciale (le visage est un « objet » structuré mais déformable) et l'orientation du visage ; mais aussi des paramètres extrinsèques comme les conditions d'acquisition de l'image (surexposition, ombres portées sur le visage …) ou les occultations partielles. Si les premières expériences se faisaient dans des conditions très contraintes (visages frontaux sur fond homogène), les contraintes ont été progressivement relâchées. Toutefois, si reconnaître une personne filmée de face semble maintenant relativement aisé ; si la personne est de profil, porte un chapeau et est filmé en contrejour par une webcam, les algorithmes actuels ont peu de chance d'aboutir [ZCR03[74]].

La détection/localisation de visage est l'étape préalable à tout système d'analyse de visage. Elle a fait l'objet de nombreux travaux, décrits dans un état de l'art très complet [YKA02[75]]. On peut regrouper ces méthodes selon deux grandes classes communes à la reconnaissance des formes : structurelle et globale. Les approches structurelles [YH94[76]] cherchent à détecter des éléments caractéristiques du visage (yeux, bouche, nez, contour de la tête) puis à combiner les résultats de ces détections grâce à des modèles géométriques et radiométriques, notamment via des modèles déformables [YHC92[77]], ou par l'analyse de "constellations" [BH03[78]]. Les approches globales sont basées sur l'apprentissage d'un modèle (génératif ou discriminant) de visage sur une base d'exemples. Elles s'appuient généralement sur un codage rétinien de l'image : une fenêtre (rétine) de tille fixe parcourt l'image avec un certain pas et la réponse du modèle est évaluée en chaque position de la rétine. L'application à des visages de tailles différentes est réalisée en itérant ce procédé, par réduction de la taille de l'image d'un facteur donné à chaque itération (processus multi-résolution).

La malédiction de la dimensionnalité (d = 400 pour une rétine 20x20) et les impératif « temps réel » des applications précédemment citées rendent la phase de génération/extraction des caractéristiques primordiale. Du codage rétinien initial (niveaux de gris) sont extraits les contours du visage et des caractéristiques faciales ; contours qui sont, a priori, relativement stables. Pour ce faire, des opérateurs « bas niveau » (gradient, orientations …) sont parfois mis en œuvre. Dans la même veine, les filtres de Haar ont été popularisés par le détecteur de

[74] Zhao W., Chellappa R., Rosenfeld A. & Phillips P.J., Face Recognition: A Literature Survey, ACM Computing Surveys, pp. 399-458, 2003.
[75] Yang M.H., Kriegman D. & Ahuja N., Detecting Faces in Images: A Survey, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 1, pp. 34-58, 2002.
[76] Yang G. & Huang T. S., Human Face Detection in Complex Background, Pattern Recognition, 27(1), pp. 53-63, 1994.
[77] Yuille A., Hallinan P. & Cohen D., Feature Extraction from Faces Using Deformable Templates, International Journal Computer Vision, 8(2), pp. 99-111, 1992.
[78] Bileschi S.M. & Heisele B., Advances in Component Based Face Detection, IEEE International Workshop on Analysis and Modeling of Face and Gestures, 2003.

Viola et Jones [VJ01][79]. La caractérisation des contours (hautes fréquences) et des zones homogènes (basses fréquences) peut aussi se baser sur la représentation espace-fréquence de la transformée en ondelettes [SK00[80]] et ses dérivés : les filtres de Gabor [HSK05[81]]. Enfin, les méthodes de réduction de la dimension basées sur la détermination d'un sous-espace sont très largement utilisées : analyse en composantes principales [MP97[82]], analyse discriminante [YKA01[83]] ….

Dans l'étape de classification (détection), les méthodes statistiques ont été largement employées. Mais les équipes de recherche se sont heurtées à un problème spécifique, la détection, qui est une classification à deux classes, la classe « objet » (visage, exemples positifs) et la classe « non-objet » (négatifs). Si la première peut être correctement représentée par une base d'exemple, la seconde est plus difficile à appréhender : comment définir un « non-visage » ? La procédure de *bootstrapping* itératif proposé dans [SP98[84]] a permis de gérer imparfaitement cet aspect du problème : l'apprentissage du classifieur commence avec l'ensemble des positifs et une sélection de négatifs. Un test détermine ensuite l'ensemble des faux positifs qui est ajouté à la base d'apprentissage. Une solution alternative consiste à utiliser une méthode générative, appris sur les exemples positifs au lieu de méthodes discriminantes qui nécessitent des exemples positifs et négatifs. Ainsi, la modélisation des visages par un mélange de gaussiennes, suivie d'une décision par maximum de vraisemblance a montré son efficacité [MP97]. Au contraire, dans [SP98], les classes visage et non-visage sont modélisées, et un réseau discriminant est mis en œuvre. De même, les réseaux à convolution ont été largement utilisés [RBK98[85], GD04[86]]. Quant aux méthodes génératives neuronales, elles seront décrites par la suite. Une version contrainte a été proposée dans [FBV01[87]]. Enfin, le détecteur proposé par Viola et Jones [VJ01] occupe une place particulière. Basé sur une cascade attentionnelle de classifieurs (filtres de Haar) *boostés*, c'est le premier détecteur ayant à la fois des performances comparables à l'état de l'art, des capacités temps réel et une généricité suffisante pour permettre son application à d'autres types d'objet. Les variantes et améliorations de cet algorithme n'ont cessé de fleurir ces dernières années : extension des caractéristiques utilisées [LM02], modification dans la construction des classifieurs forts [HAW04[88], LZ04[89]], structure arborescente [KKK05[90]] pour détecter les visages quelque soit leur orientation …

---

[79] Viola P. & Jones M., Rapid object detection using a boosted cascade of simple features, International Conference on Computer Vision and Pattern Recognition, (1), pp 511-518, 2001.

[80] Schneiderman & Kanade T., A statistical model for 3D object detection applied to faces and cars, Conference on Computer Vision and Pattern Recognition, (1), pp 746-751, 2000.

[81] Huang L. L., Shimizu A. & Kobakate H., Robust face detection using Gabor filter features, Pattern Recognition Letters, 26(11):1641-1649, 2005.

[82] Moghaddam, B. Pentland, A.: Probabilistic visual learning for object representation, IEEE Transactions on Pattern Analysis and Machine Intelligence, 19(7), 696-710, 1997.

[83] Yang M.H., Kriegman D. & Ahuja N., Face detection using multimodal density models, Computer Vision and Image Understanding, 84(2), pp. 264-284, 2001.

[84] Sung K.K. & Poggio T., Example-based learning for view-based human face detection, IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(1), pp 39-51, 1998.

[85] Rowley H.A., Baluja S. & Kanade T., Neural network based face detection, IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(1): 23-38, 1998.

[86] Garcia C. & Delakis M. (2004) Convolutional face finder: A neural architecture for fast and robust face detection, IEEE Transactions on Pattern Analysis and Machine Intelligence, 26(11): 1408-1423.

[87] Féraud R., Bernier O., Viallet J., Collobert M.: A fast and accurate face detector based on neural networks, IEEE Transactions on Pattern Analysis and Machine Intelligence, 23(1), 42-53, 2002.

[88] Huang C., Ai H., Wu B. & Lao S., Boosting nested cascade detector for multi-view face detection, International Conference on Pattern Recognition, (2), pp 415-418, 2004.

[89] Li S.Z. & Zhang Z., Floatboost learning and statistical face detection, IEEE Transactions on Pattern Analysis and Machine Intelligence, 26(9):1112-1123, 2004.

En vue d'applications similaires, de nombreuses méthodes de détection et localisation des caractéristiques faciales ont été développées ces dernières années. Toutefois, la grande majorité d'entre elles se concentrent sur la détection des yeux qui sont la caractéristique la plus intéressante (pour la biométrie par l'iris, en particulier) mais aussi la plus aisément détectable (l'apparence du nez change avec la pose du visage et celle de la bouche avec l'expression faciale). Les méthodes de détection des caractéristiques faciales peuvent être divisées en deux catégories. Les méthodes de traitement d'images mettent en œuvre un ou plusieurs détecteurs « bas niveau » (contour, symétrie, couleur …) et combinent leur sortie [FY01[91]]. Les méthodes basées sur l'apparence s'inspirent des algorithmes d'apprentissage statique ou dynamique (actif) déjà utilisés en détection de visages. Une analyse en composantes principales locale est utilisée dans [MP97] pour décrire les caractéristiques faciales dans un sous-espace appelé espace des caractéristiques propres (*eigenfeature space*). Dans [DF05[92]], un réseau de neurones à convolution complexe est proposé. Le détecteur décrit dans [CC03[93]] est une adaptation de celui proposé dans [VJ01]. Ils démontrent que la performance de ces détecteurs locaux peut être améliorée en ajoutant des contraintes plus globales sur les positions relatives des hypothèses de détection les unes par rapport aux autres. Les méthodes actives sont aussi largement employées : modèles déformables [YHC92] et, plus récemment, modèles d'apparence actifs [CET98[94]]. Ces dernières méthodes, quoiqu'élégantes, sont aussi gourmandes en temps de calcul et particulièrement sensibles au bruit, aux occultations et aux conditions initiales.

3.3.2 Localisation de visages dans une image [Annexe 4]

Co-encadrements : 1 thèse (Belaroussi R.), 1 stage (Torkamanlou P.)

3.3.2.1 Principe

Comme précédemment, nous avons souhaité résoudre ce problème en le décomposant en sous tâches plus aisément réalisables. Au lieu de détecter directement le visage dans l'image, nous avons utilisé notre bon sens pour définir ce qu'était un visage. La définition que nous avons retenue s'énonce comme suit « objet de forme elliptique dont les composantes (yeux, bouches …) sont organisées spatialement et dont la teinte est particulière ». Nous cherchons donc dans la suite à caractériser les visages par leurs propriétés géométrique, anthropomorphique et colorimétrique. La combinaison de ces trois informations est d'autant plus pertinente que certaines d'entre elles peuvent disparaître (occultations) ou n'être que peu discriminantes (multiplicité des teintes « chair »). Les expérimentations ont été conduites sur la base ECU [PBC05[95]] qui contient plus de 3000 images couleur étiquetées (rectangle englobant le visage et zones de teinte chair). La tâche de localisation sur cette base est particulièrement complexe en raison de la grande variabilité apparaissant dans les images (Figure 18) au niveau des caractéristiques intrinsèques (origine ethnique, âge, orientation) et extrinsèques (taille dans

[90] Kim J.B., Kee S.C. & Kim J.Y., Fast detection of multiview face and eye based on cascaded classifier, Machine vision and Applications, 116-119, 2005.
[91] Feng G.C. Yuen P.C., Multi-cues eye detection on gray intensity image, Pattern Recognition, 34, 1033-1046, 2001.
[92] Duffner S., Garcia C., A Connexionist Approach for Robust and Precise Facial Feature Detection in Complex Scenes, IEEE International Symposium on Image and Signal Processing and Analysis, 316-321, 2005.
[93] Cristinacce D., Cootes T.: A comparison of shape constrained facial feature detectors, International Conference on Automatic Face and Gesture Recognition, 375-380, 2004.
[94] Cootes T.F., Edwards G.J. & Taylor J.C., Active Appreance Models,. European Conference on Computer Vision, pp 484-498, 1998.
[95] Phung S.L., Bouzerdoum A. & Chai D., Skin segmentation using color pixel classification: Analysis and comparison, IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(1), pp 148-154, 2005.

l'image, résolution du capteur). Les données ont été distribuées en deux corpus de même taille, le premier incluant les bases d'apprentissage et de validation croisée, le second correspondant à la base de test.

3.3.2.2 Modèle d'apparence du visage : réseau diabolo

Afin de nous libérer de la contrainte de « trouver » des exemples négatifs (non-visages), nous avons opté pour un modèle d'apparence génératif précédemment utilisé par l'équipe en reconnaissance de caractères : un réseau de neurones auto-associatif (réseau Diabolo [SM96[96]]). Ce réseau est entraîné à reconstruire en sortie une image identique à celle d'entrée en réalisant une compression spécialisée (analyse en composante principale non linéaire [BH89[97]]) car la couche cachée comporte un nombre de cellules nettement inférieur à celui de l'entrée ou de la sortie. Par analogie avec l'analyse en composante principale, en augmentant le nombre de neurones sur la couche cachée, on capture une part croissante de l'inertie des données, améliorant d'autant les capacités de modélisation du réseau. Au-delà d'une certaine valeur, l'erreur en apprentissage ne décroît plus significativement, les dimensions ajoutées ayant une très faible inertie (bruit). En test, une image de non-visage sera en principe mal compressée et donnera une erreur de reconstruction bien plus importante qu'une image de visage (Figure 1). Les poids du réseau de neurones sont estimés par rétro-propagation avec adaptation du pas d'apprentissage et arrêt par validation croisée. Une fois le réseau entraîné, une image peut être traitée en appliquant la procédure suivante. L'image en niveau de gris est parcourue, à l'échelle correspondant à la taille du visage, par une fenêtre glissante et, en chaque pixel, une erreur de reconstruction est calculée. Une fois balayée toute l'image, on obtient une carte des erreurs de reconstruction (notée D : Figure 14) dont le maximum nous informe sur la localisation du visage dans l'image.

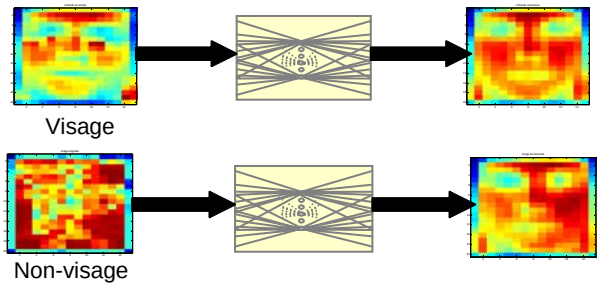


Figure 13. Reconstruction d'un visage et d'un « non-visage » par réseau diabolo.

Une fois choisi le modèle d'apparence, reste à sélectionner un espace de représentation convenable. Or, une partie de l'information contenue dans l'image d'un visage se trouve dans l'orientation du gradient des contours. Le premier avantage des contours sur les niveaux de gris est leur relative invariance à la couleur de la peau. Le second est lié à la présence de parties concaves (yeux, bouche) et convexes (nez) dans un visage qui créent des contours relativement stables. Chaque pixel (i,j) de contour de la vignette est donc représenté par deux caractéristiques correspondant aux sinus et cosinus de l'orientation du gradient. Les pixels hors contours reçoivent les valeurs (0,0). Les deux images d'orientation sont filtrées à l'aide d'un masque elliptique afin d'éliminer le fond et fusionnées pour former le vecteur d'entrée du réseau. Nous avons ainsi montré que les performances [d'un localiseur traitant] des images

[96] Schwenk H. & Milgram M., Constraint tangent distance for on-line character recognition, International Conference on Pattern Recognition, (D), pp 520-524, 1996.
[97] Baldi P. & Hornik K., Neural networks and principal component analysis: learning from examples without local minima, Neural Network, 2(1):53-58, 1989.

d’orientation de gradient étaient meilleures que celles des images de gradient ou de luminance.

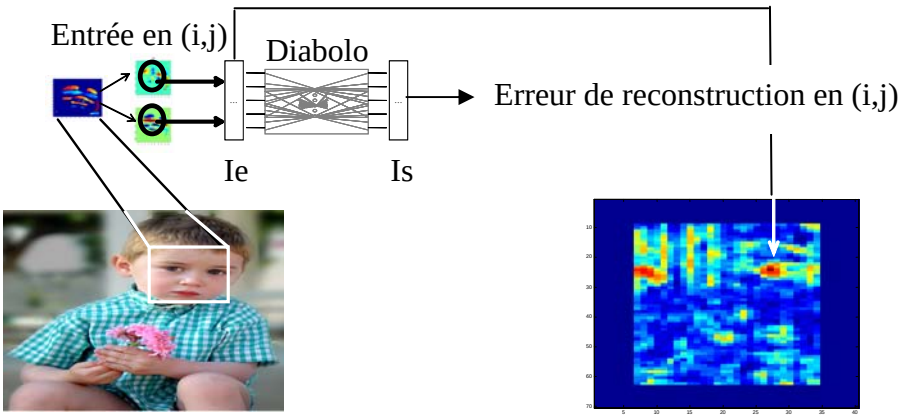
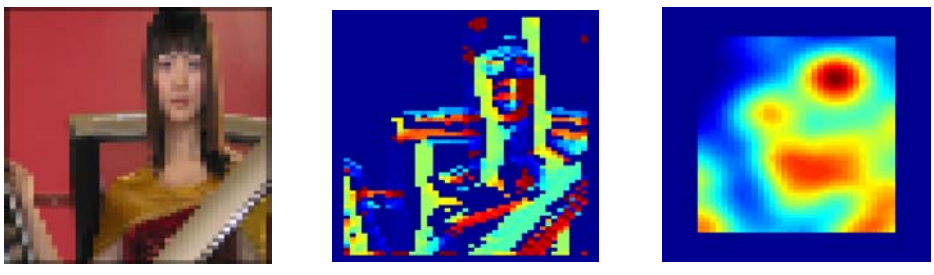


Figure 14. Carte des erreurs de reconstruction du diabolo

3.3.2.3 Modèle géométrique : détecteur d’ellipse basé sur la Transformation de Hough Généralisée

La forme elliptique du visage est détectée par Transformation de Hough Généralisée (THG) : les visages sont modélisés par une ellipse verticale d’excentricité donnée. Après normalisation de l’image de luminance, l’orientation du gradient est estimée et quantifiée sur N valeurs. Puis, les point de contour sont déterminés par seuillage du module du gradient (Figure 15.b). La THG est alors appliquée dans sa forme simplifiée, basée sur les propriétés géométriques des ellipses. La méthode consiste à accumuler les votes de tous les pixels d’un segment partant d’un point de contour et dont la direction et le sens sont donnés par l’orientation du contour en ce point. Chaque pixel incrémente les valeurs d’un tableau de vote : l’accumulateur. La position du maximum de l’accumulateur correspond à la position dans l’image du point le plus susceptible d’être le centre de l’ellipse. Comme le fond est souvent structuré et complexe, ce maximum ne localise qu’approximativement la position du visage. Afin de diminuer l’effet du fond sur le tableau de vote, ce dernier est d’abord lissé puis parcouru par un filtre en "chapeau mexicain" afin de détecter les maximas locaux. Là encore, le visage est localisé au maximum de la carte résultante, qui est le maximum global de l’accumulateur (notée H : Figure 15.c).



(a) (b) (c)

Figure 15. Image originale (a), orientation du gradient (b) et tableau de vote H (c).

3.3.2.4 Modèle colorimétrique : détecteur de teinte chair

Contrairement aux deux experts précédents qui caractérisaient respectivement l'apparence et la forme d'une région de l'image, la détection de la teinte chair cherche à déterminer si un pixel appartient ou non à la peau d'un individu présent dans l'image. Les principales difficultés rencontrées par une telle classification sont la présence de pixels de fond ayant une couleur voisine de celle de la peau, la multiplicité des sources lumineuses entraînant l'apparition de zones d'ombre et de surbrillance et, bien sur, les différentes teintes possibles de la peau fonction de l'origine ethnique du sujet. Plusieurs études ont cependant démontré que la couleur de peau était en majeure partie confinée dans un volume restreint de l'espace [BM01[98]], et ce, dans plusieurs espaces colorimétriques. Parmi ces derniers, l'analyse de l'existant a montré la supériorité des espaces YCbCr et HSV, dont le principal avantage est de séparer l'information de luminance (Y ou V) de celle de chrominance (CbCr et HS). Nous avons ensuite évalué plusieurs détecteurs : définition explicite de frontières dans l'espace colorimétrique, modélisation non paramétrique (histogramme) et paramétrique (avec dépendance fonctionnelle gaussienne). Nous listons dans la suite les principales conclusions de cette étude :

- les modèles non paramétriques sont plus performants que les modèles paramétriques (gaussien ou mélange de gaussiennes), conclusion assez logique puisque les premiers ne font aucune hypothèse sur la forme des densités de probabilité.
- dans l'espace CbCr, la modélisation de la peau et de la non-peau augmente le taux de détection comparativement à la modélisation de la peau seule. Dans l'espace HS, au contraire, cette dernière suffit, évitant la recherche d'exemples négatifs.
- l'information de luminance améliore aussi le taux de détection : la partition de cette dernière en plusieurs intervalles permet de mieux estimer les densités de probabilité.

L'image de détection est finalement moyennée. Nous noterons S la carte résultante.

3.3.1.5 Fusion d'experts pour la localisation de visage

Nous utilisons l'information produite par les trois détecteurs précédents : modèle géométrique, modèle d'apparence et modèle de teinte chair. Ces détecteurs produisent, pour une image couleur donnée, trois cartes D, H et S. La combinaison de ces trois cartes (Figure 16) permet d'effectuer la localisation lorsqu'un des experts est insuffisant ou que les trois experts sont en conflit. La carte de probabilité résultante de la combinaison est notée F.

Les trois cartes D, H et S sont normalisées dans l'intervalle [-1 ;1]. Une rétine centrée sur le pixel (i,j) dans l'image originale est alors caractérisée par le vecteur de descripteurs $[D_{i,j}\ H_{i,j}\ S_{i,j}]$. Plusieurs algorithmes de fusion d'informations paramétrique ont été évalués : bayésien, flou et neuronal. Tous produisent une carte F = f(D,H,S) dont le maximum indique la position du visage dans l'image. Le tableau 1 permet de comparer les performances des détecteurs initiaux et des algorithmes de fusion. Comme on peut le constater, la fusion neuronale obtient les meilleurs résultats ; et ce avec un réseau de neurones intégrateur réduit à sa plus simple expression : un neurone à transfert linéaire unique ! Ce dernier réalise donc une somme pondérée des sorties des trois détecteurs suivant l'équation : $F_{i,j} = [a\ b\ c][D_{i,j}\ H_{i,j}\ S_{i,j}]^T$. En comparaison, la méthode de fusion non paramétrique basée sur la somme des sorties des détecteurs est déjà assez efficace.

[98] Brand J. & Mason J.S., Skin probability map and its use in face detection, International Conference on Image Processing, (1), 1034-1037, 2001.

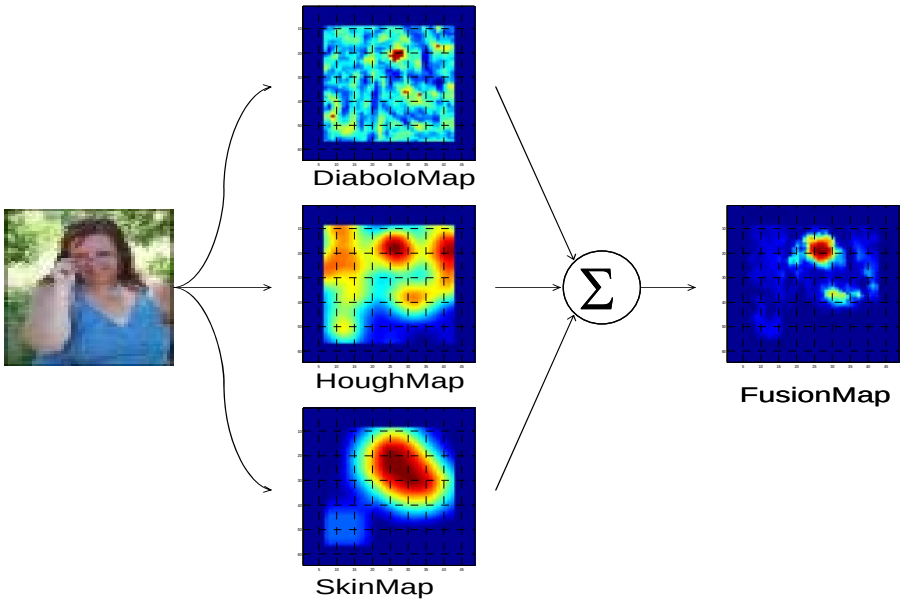


Figure 16. Système de localisation de visage.

Localiseur	Modèle d'apparence	Modèle d'ellipse	Flou	Bayésien	Neuronal	Somme
Taux de localisation	49%	67%	72%	84%	86%	80%

Tableau 9. Taux de localisation des différentes approches (Base ECU – Test : 1353 images).

Les précédents résultats amènent une question quant à la mesure de performance de performance utilisée. Il est en effet difficile de définir une métrique permettant de juger de la précision d'une localisation. Nombre de communications du domaine omettent purement et simplement cet aspect, pourtant crucial. De fait, il n'existe pas de métrique « universelle » mais presqu'autant de mesures que d'équipes de recherche [RCB06[99]]. Nous avons choisi de considérer un visage correctement localisé lorsque le recouvrement entre vérité terrain et localisation était supérieur à 60%. La Figure 17 montre la répartition de ces taux de recouvrements sur la base de test. Concentrons-nous sur les deux extrémités de l'histogramme. A gauche, les modèles géométrique et d'apparence génèrent beaucoup de fausses alarmes dues au fond complexe. Comme ces erreurs apparaissant rarement pour au même endroit, les conflits entre ces deux sources sont nombreux. La fusion parvient à les éliminer, diminuant d'autant les fausses alarmes. A droite, la fusion réussit à augmenter le nombre de visages bien localisés, en s'appuyant sur l'information colorimétrique..

Un premier test a été effectué sur les 1353 images de la base ECU ne contenant qu'un visage, à une seule échelle, c'est-à-dire en supposant connue la distance séparant la caméra u sujet (Figure 18.a). Un second test est réalisé sur les 205 images restantes, totalisant 482 visages. Dans les images "mono-visage" la position du visage est définie par celle du maximum de F. Dans une image contenant N visages (N étant supposée connu pour un problème de localisation) les N plus grand maximas locaux (suffisamment distants pour éviter les recouvrements) définissent la position des visages. 396 visages sont correctement détectés (Figure 18.b) sur 482 (82%). Enfin, nous avons souhaité évaluer les capacités du système à localiser un visage de taille quelconque. Pour ce faire, une pyramide d'images est construite : l'image initiale est réduite d'un facteur (généralement fixé à la valeur 1.2) et parcouru par une

[99] Rodriguez Y., Cardinaux F., Bengio S. & Mariethoz J., Measuring the performance of face localization systems, Image and Vision Computing, 24(8): 882:893, 2006.

rétine de taille fixe. Les résultats obtenus alors ont été un peu décevants avec un taux de localisation de 58%. Le détecteur de Viola & Jones (dont la première version est disponible sous OpenCv), testé dans les mêmes conditions, a quant à lui détecté 71% des visages et produit 50 fausses alarmes. Toutefois, l'analyse qualitative des hypothèses émises par les deux systèmes ont montré une certaine complémentarité. Le détecteur de viola & Jones fournit généralement une localisation plus précise, mais échoue en présence de visages à l'orientation trop marquée ou partiellement occultés.

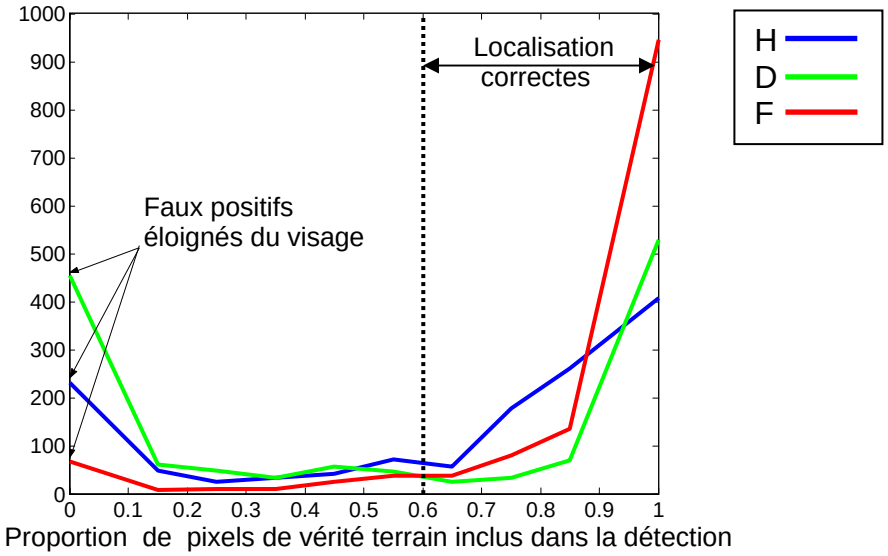
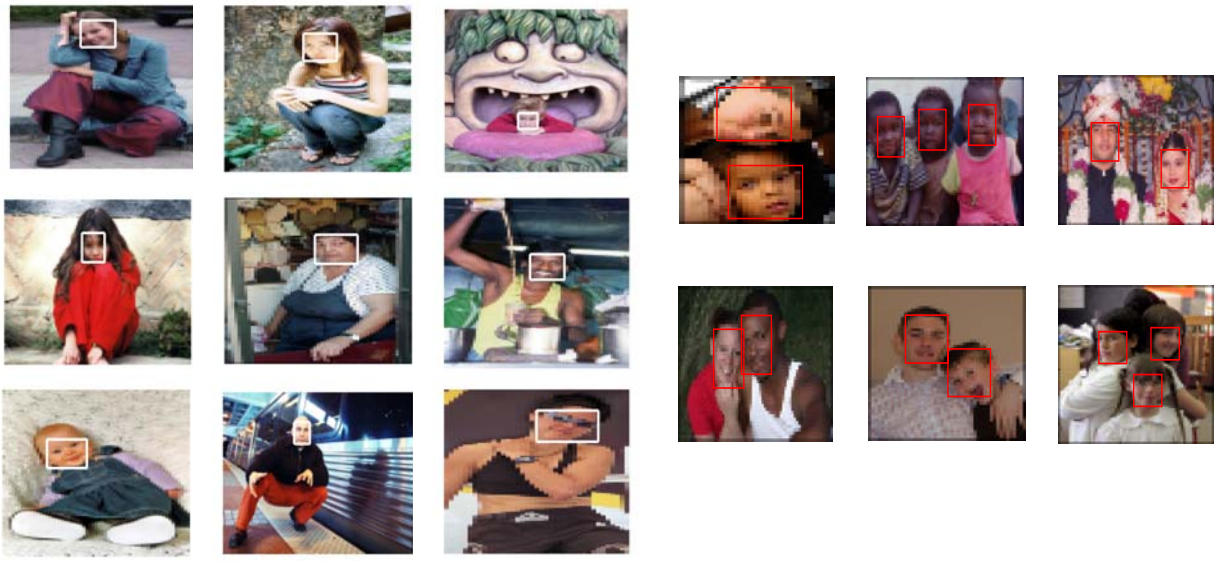


Figure 17. Nombre de visages en fonction du taux de recouvrement entre vérité terrain et localisation



(a) (b)

Figure 18. Exemples de localisation mono-visage (a) et multi-visages (b).

3.3.3 Localisation des caractéristiques faciales [Annexe 5]

Encadrements : 2 stages (Eshkevari B., Muhamad Hanif S.)

3.3.3.1 Principe

Le réseau auto-associatif hybride est un perceptron à deux couches complètement connectées. Contrairement au réseau auto-associatif classique (diabolo) qui apprend à reconstruire son entrée, ce réseau hybride est entraîné à associer une image de visage (entrée) à une carte de caractéristiques (sortie) en vue de détecter ces caractéristiques. La couche cachée, de petite taille, réalise une compression (analyse en composante principale non linéaire) des entrées. La fonction de coût est l'erreur quadratique moyenne entre la sortie de réseau et la sortie désirée (Figure 19). L'apprentissage du réseau est fait par rétro-propagation du gradient de l'erreur. Une fois le détecteur entraîné, la détection des caractéristiques faciales est réalisée en cherchant les maxima locaux dans la sortie de réseau et en les rétro-projetant sur l'image originale (Figure 20).

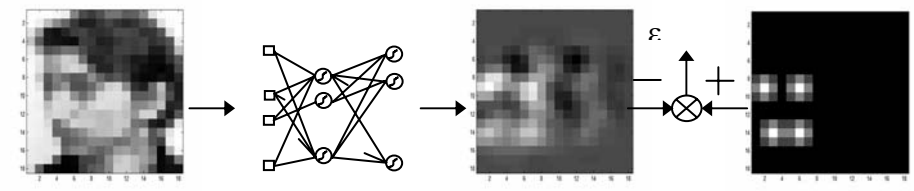


Figure 19. Apprentissage : image en entrée de réseau(a). La fonction de cout est l'erreur quadratique moyenne entre sortie du réseau (b) et carte de caractéristiques (c).

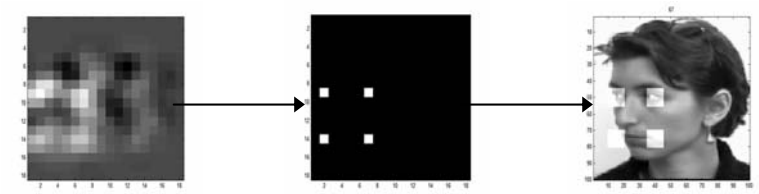


Figure 20. Décision : sortie du réseau (a), 4 premiers maxima locaux(b), projection dans l'image originale (c).

Nous avons utilisé deux bases d'images : la base ECU décrite précédemment et la base LISIF. Cette dernière contient environ 1400 images de visage de 37 personnes de différents âges, sexes et origines ethniques prises dans des conditions contraintes (éclairage, fond) avec différentes poses (suivant les trois angles d'orientation possibles du visage), expressions faciales et « accessoires » comme barbe et lunettes. Ces bases ont été annotées manuellement et quatre caractéristiques faciales (yeux et commissures des lèvres) ont été marquées pour construire une carte de caractéristiques *F*. Cette carte a la même taille que l'image de visage et ses pixels ont les valeurs suivantes : $F(x_{oi},y_{oi}) = +1$ au niveau des caractéristiques et $F(i,j) = -1$ ailleurs ; avec (x_{oi},y_{oi}) les coordonnées de la $i^{ème}$ caractéristique.

3.3.4.2 Variantes

Nous proposons trois variantes au localiseur précédent, en vue d'améliorer sa précision et sa robustesse : la première est un réseau de neurones à décalage spatial partiellement connecté, la seconde est une combinaison parallèle de plusieurs réseaux spécialisés, enfin, nous présentons une cascade à deux niveaux capable d'exécuter la tâche en temps réel.

Les réseaux de neurones à convolution – appelés réseaux de neurones à décalage spatial (*Space Displacement Neural Networks* - SDNN) en analyse d'image – sont un peu différents des réseaux complètement connectés. Contrairement aux MLP classiques où un neurone d'une couche est connecté à tous les neurones de la couche précédente, les réseaux de neurones à convolution possèdent des zones réceptives locales. Un neurone n'est connecté qu'à un sous-ensemble de neurones de la couche précédente. Chaque neurone peut être vu comme une unité de détection d'une caractéristique locale. Le concept de zone réceptive est issu d'expériences faites en psychologie de la perception [HW62[100]]. Le réseau de neurones à décalage spatial (Figure 21) ne comporte qu'une seule couche d'extraction de caractéristiques. Les deuxième et troisième couches ont le même rôle associatif que dans l'architecture précédente. Plusieurs couches d'extraction de caractéristiques peuvent bien sur être ajoutées pour augmenter la capacité de représentation du réseau [LBB98[101]].

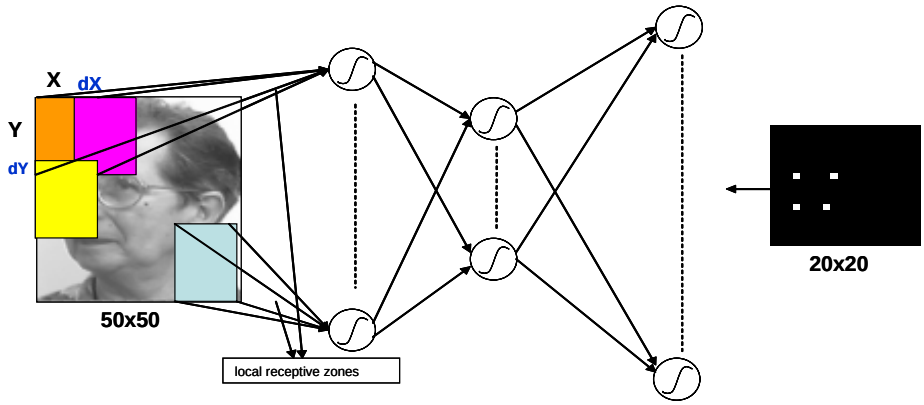
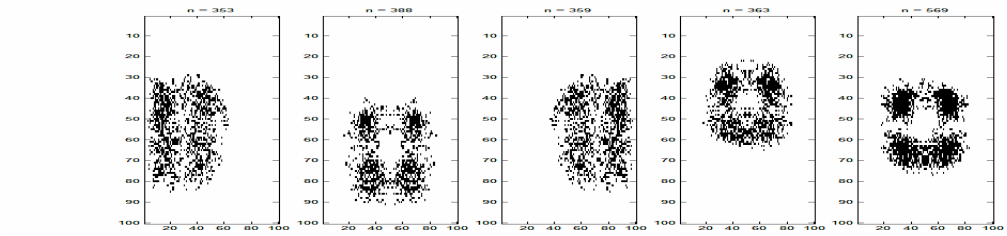


Figure 21. Réseau auto-associatif à connexions locales

Pour améliorer la précision du réseau localiseur, nous avons décidé d'utiliser plusieurs réseaux ; chacun spécialisé sur une orientation donnée. Le mélange d'orientations peut être modélisé par un mélange de gaussiennes, une pour chaque orientation. Dans ce dernier cas, nous estimons les paramètres sur un cluster. Nous employons l'algorithme EM, initialisé en utilisant l'algorithme K-*means* et adapté sur 1000 itérations. Nous avons considéré jusqu'à six orientations et voyons que, pour cinq clusters (Figure 22), le *clustering* a grossièrement séparé la base en sous-ensembles, chacune représentant une certaine orientation.



(a) (b) (c) (d) (e)

Figure 22. Position des caractéristiques faciales selon l'orientation du visage : vers la gauche (a), frontale vers le bas (b), vers la droite (c), frontale vers le haut (d), frontale (e).

[100] Hubel D. & Wiesel T. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex, Journal of Psychology, 160: 106-154.

[101] LeCun Y., Bottou L., Bengio Y. & Haffner P. (1998) Gradient Based Learning Applied to Document Recognition, Proceedings of IEEE, 86(11): 2278-2324.

En considérant N orientations, le localiseur multiple correspondant comporte N réseaux, chacun appris sur un sous-ensemble et produisant une carte de caractéristiques. Nous avons employé un réseau intégrateur pour combiner ces hypothèses. Il s'agit d'un cas particuliers des ensembles de réseaux (Figure 23) qui sont des outils particulièrement puissants pour faire face à des problèmes complexes. Ils se composent d'une combinaison linéaire de plusieurs réseaux (experts) qui ont été entraînés en utilisant les mêmes données. Chaque réseau dans l'ensemble a un poids potentiellement différent dans la sortie de l'ensemble. Ce poids est appris à l'aide d'un réseau de neurone supplémentaire que nous appelons réseau intégrateur puisqu'il produit une nouvelle carte de caractéristiques, somme pondérée des N sorties. Ces poids peuvent aussi être « binarisés » en conservant le poids le plus grand et en annulant tous les autres. On a alors un réseau « porte » (*gating network*) qui filtre les hypothèses et ne conserve que la plus plausible. L'ensemble de réseaux a une erreur en généralisation et une variance plus faibles que celles obtenues avec un réseau simple [PC93[102]]. Ici, contrairement au cas général, les experts et le réseau intégrateur sont entraînés séparément. Dans un premier temps, chaque expert est entraîné sur une orientation de visage. Puis les poids du réseau intégrateur sont calculés à l'aide de l'algorithme « *Generalized Ensemble Method* » (GEM). L'algorithme GEM calcule la déviation entre la sortie de chaque expert et la sortie désirée et, en utilisant la matrice de corrélation, calcule les poids de la combinaison. Le réseau intégrateur est un Perceptron bi-couche complètement connecté. L'entrée du réseau est une image de visage et ses sorties sont les coefficients de la combinaison.

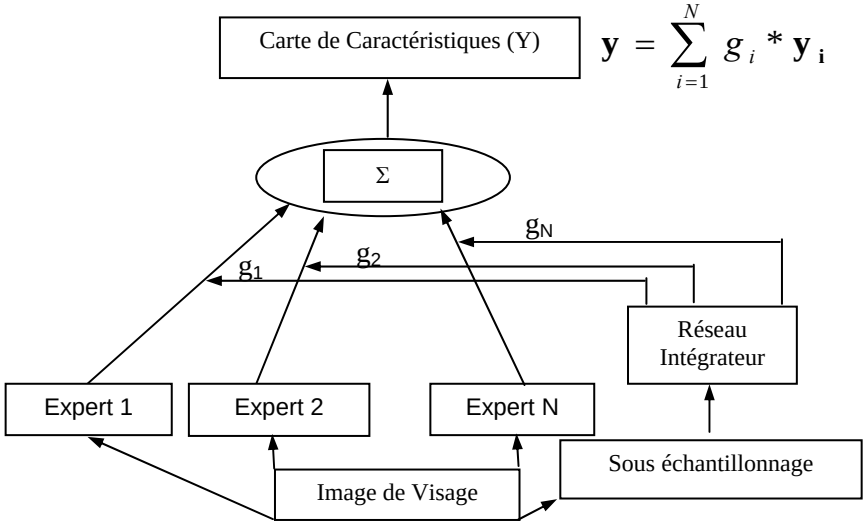


Figure 23. Localiseur multiple : ensemble de réseaux et réseau intégrateur.

3.3.4.3 Résultats expérimentaux

La mesure de performance, est l'erreur de localisation normalisée EL i.e. la somme des distances euclidiennes entre les caractéristiques détectées et la vérité terrain divisée par la distance interoculaire (distance euclidienne entre l'œil gauche et l'œil droit). L'erreur de localisation moyenne (ELM) est ensuite calculée sur toutes les images de test. Tous les algorithmes présentés ont été implémentés sur un Intel Pentium Centrino (1.6 GHz) sous MATLAB.

[102] Perrone M.P. & Cooper L. N., When Network Disagrees: Ensemble Methods for Hybrid Neural Network, Neural Networks for Speech and Image Processing, Chapman - Hall, pp 126-142, 1993

Localiseur simple (SMLP)

Nous avons d'abord entraîné un unique réseau pour la localisation des caractéristiques de visage sur la base entière indépendamment de l'orientation. Les résultats obtenus montrent une relative insensibilité au nombre de caractéristiques à détecter (1, 2 ou 4). Pour évaluer la sensibilité à l'identité, nous avons adopté une stratégie proche du « *leave-one-out* » dans laquelle nous testons les réseaux sur les images d'une personne et utilisons les images de toutes les autres personnes pour l'apprentissage. Les résultats obtenus (Tableau 9) sont décevants, quoique prévisibles. La moyenne et la variance de l'erreur de localisation sont élevées ce qui peut s'expliquer par le faible nombre d'individus présents dans la base d'apprentissage. Le localiseur apprend les caractéristiques anthropomorphiques de ces individus et est incapable de généraliser le processus d'association entre image de visage et carte de caractéristiques à d'autres individus. Au contraire, si tous les individus sont présents dans les bases d'apprentissage et de test, l'association se fait correctement et l'erreur moyenne de localisation chute. Dans les deux cas, la médiane de l'erreur de localisation est significativement plus faible que la moyenne, ce qui montre que le localiseur échoue pour certaines orientations trop marquées, probablement absentes de la base d'apprentissage.

Statistiques	**ELM**	**ELM « *leave-one-out* »**
Moyenne	0,16	0,28
Ecart Type	0,10	0,16
Médiane	0,12	0,24

Tableau 10. Sensibilité à l'orientation du visage et à l'identité de personne – Localiseur : SMLP, base : LISIF.

Localiseur Multiple (MMLP)

Le clustering décrit dans la section 2 a séparé l'ensemble de données initial en plusieurs sous-ensembles correspondant chacun à une pose du visage. Pour une image d'entrée, nous avons N images de sortie et N hypothèses de localisation contenant les quatre maxima de chaque image de sortie. Pour évaluer la précision du localiseur multiple, nous calculons l'erreur de localisation moyenne pour chaque hypothèse et appliquons le critère WTA (*Winner Takes All*) pour sélection la meilleure (Figure 24 et Tableau 11). Nous observons que l'erreur moyenne de localisation décroît continûment quand N augmente, sur la base d'apprentissage comme sur la base de test. Ces résultats sont tout à fait logiques : lorsque que le nombre de réseaux spécialisés augmente, la gamme des orientations de visage que chaque réseau doit traiter diminue. Le processus d'association entre l'image de visage et la carte de caractéristique devient plus facile et l'erreur diminue.

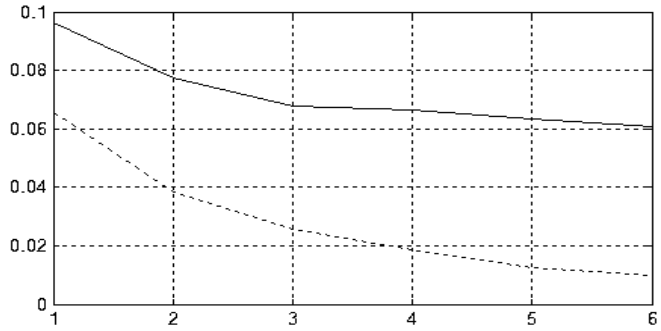


Figure 24. Influence du nombre d'orientations sur l'ELM en apprentissage (trait pointillé) et en test (trait plein) – localiseur : MMLP, base : LISIF.

Localiseur Multiple (MSDNN)

En conservant les paramètres précédents, nous avons entraîné cinq SDNN spécialisés sur chaque sous-ensemble d'image. Nous avons également entraîné un réseau intégrateur à 80 neurones cachés (Figure 25). Les performances du localiseur multiple sur la base de test sont les suivantes : l'ELM est de 0,12 et 65% des exemples ont une erreur de localisation inférieure à 0,1 (Tableau 11). Le système (réseaux multiples et réseau intégrateur) peut traiter 11 images/seconde.

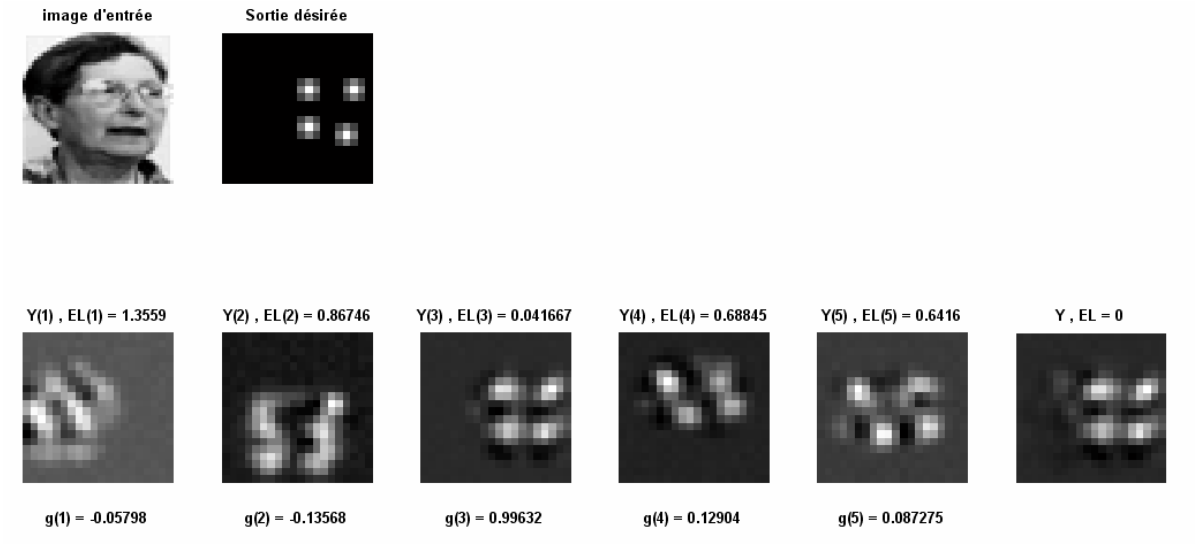


Figure 25. Localiseur multiple (MSDNN), Y(i), g(i), EL(i) pour i = 1, 2, … , 5 sont la sortie, le poids du réseau intégrateur et l'erreur de localisation de chaque expert. Y est la sortie du localiseur multiple et EL l'erreur de localisation pour l'exemple. Nous voyons que l'erreur de l'ensemble est inférieure à celles des experts.

Nous présentons (Figure 26) quelques résultats de localisation sur des images de test de la base. Comme on peut le voir, le localiseur est robuste quelque soit l'orientation du visage et en présence d'accessoires (barbe, lunettes). Le Tableau 11 contient une synthèse des résultats de localisation sur les deux bases. Nous constatons (lignes 1 et 3) que le SDNN généralise mieux que le MLP lorsque le nombre d'exemples d'apprentissage augmente. Le problème de sensibilité à l'identité de la personne - point faible du localiseur entraîné sur la base LISIF – a disparu. Enfin (ligne 4), la spécialisation des localiseurs à un type d'orientation et la fusion par réseau intégrateur améliore la précision : l'ELM baisse à 0,12. Nous avons aussi évalué, sur des images synthétiques, la sensibilité des différentes variantes du localiseur aux occultations et au bruit. Là encore, le MSDNN s'avère particulièrement robuste.

Finalement, pour remplir la contrainte temps réel, nous proposons de mettre en cascade les localiseurs simple et multiple. Le localiseur simple peut traiter 110 images/secondes tandis que le localiseur multiple ne traite que 11 images/secondes. La cascade emploie le réseau simple SSDNN comme localiseur au niveau 1. Si l'hypothèse est rejetée par l'étape de validation (seuillage sur les distances de Mahalanobis inter-caractéristiques), le localiseur multiple MSDNN est activé comme localiseur au niveau 2. Si le localiseur multiple échoue, l'image est rejetée. La cascade a une erreur moyenne de 0.12 et peut traiter 40 images/seconde. Le taux de rejet en test est de 3% avec un faible pourcentage (0,6%) de détections manqués.

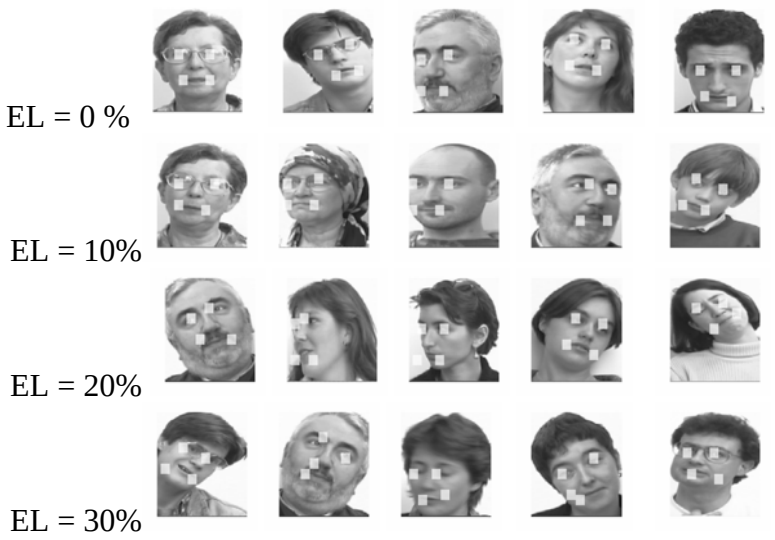


Figure 26. Résultats de la localisation sur des images de test de la base LISIF. Les images sont classées par erreur de localisation croissante– Localiseur MSDNN.

Localiseur	BASE LISIF		BASE ECU	
	Appr.	Test	Appr.	Test
SMLP	0,09	0,15	0,10	0,15
MMLP*	0,01	0,09	-	-
SSDNN	0,09	0,16	0,11	0,14
MSDNN**	0,06	0,12	-	-

*Sélection par WTA,**Fusion par Réseau Intégrateur

Tableau 11. Erreurs moyennes de localisation.

3.3.6 Conclusions et perspectives

Encadrements : 2 stages (Chetouani A. et Leroy O.), 1 projet (Ait Mohand K.)

Nous avons proposé une méthode de localisation de visages dans des images couleur basée sur la fusion des réponses de trois experts délivrant respectivement des informations géométrique, anthropomorphique et colorimétrique. Ces dernières, lorsqu'elles sont combinées de manière correcte, permettent dans la plupart des cas de localiser correctement le visage lorsque la distance séparant ce dernier de la caméra est connue. Si cette contrainte est relâchée, les performances se détériorent et les temps de calcul explosent car la recherche s'effectue à plusieurs échelles. Toutefois, l'approche a été appliquée brutalement, alors qu'un filtrage préalable des régions d'intérêt de l'image est souvent réalisé, réduisant simultanément les fausses alarmes et le temps de calcul. Le détecteur de Viola & Jones pourrait jouer ce rôle de filtrage. L'approche par fusion d'experts remplacerait alors avantageusement les derniers étages de la cascade attentionnelle, responsables du rejet des visages non frontaux ou occultés. Comme l'ont montré les expériences réalisées avec le réseau auto-associatif hybride en localisation des caractéristiques faciales, une spécialisation des modèles d'apparence et géométrique à une classe d'orientation donnée, ainsi qu'une architecture de type SDNN, permettraient d'améliorer la précision de la localisation.

Nous avons présenté un nouvel algorithme pour la localisation des caractéristiques faciales dans une image de visage. Il est basé sur un réseau de neurones particulier capable d'associer une carte de caractéristiques à une image de visage. Nous avons proposé une extension de la méthode combinant plusieurs localiseurs, chacun spécialisé sur une orientation donnée. La meilleure hypothèse de localisation est sélectionnée en combinant toutes les sorties à l'aide d'un réseau intégrateur. Ce localiseur multiple est plus précis que le localiseur unique : l'erreur de localisation moyenne diminue et le système peut traiter plus de 40 images/seconde. Nos travaux les plus récents sur la localisation des caractéristiques faciales ont montré qu'une réduction de la dimension de l'espace de représentation par analyse en composante principale sur les images de luminance conduisait à une amélioration des performances. Nous avons aussi vérifié les capacités d'interpolation et d'extrapolation du localiseur en le testant sur la base ICPR pointing'04[103] étiquetée en orientation. Ainsi, l'apprentissage du localiseur sur un intervalle limité d'orientations, suivi d'un test sur la totalité de ces dernières, ont montré que l'erreur de localisation ne variait pas significativement, même sur les orientations absentes de l'ensemble d'apprentissage. Enfin, une étape de localisation fine des caractéristiques faciales a été développée, en cascade après l'étape de localisation grossière (Figure 27). Cette localisation prendrait tout son sens dans des applications de suivi des mouvements labiaux ou pour focaliser un système de reconnaissance de l'iris [K07[104]]

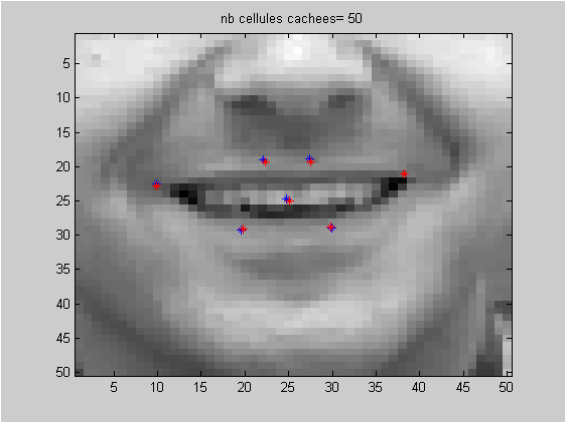


Figure 27. Localisation fine des caractéristiques faciales par réseau auto-associatif hybride.

[103] http://www-prima.inrialpes.fr/Pointing04/

[104] Krichen E., Reconnaissance de personnes par l'iris en mode degrade, Thèse de doctorat de l'Institut National des Télécommunications, 2007.

4. PROJETS EN COURS

4.1 DETECTION DE VEHICULES

Encadrements : 1 thèse (P. Negri), 1 stage (A. Belaidi)

Ces recherches ont commencé courant 2006, dans le cadre d'un contrat avec Peugeot-Citroen Automobiles (PSA). L'équipe a acquis une expertise dans le domaine des systèmes de transport intelligents à travers trois projets industriels portant respectivement sur la détection d'obstacles coopératifs (programme ARCOS), l'estimation de la courbure de la route (VALEO), et la reconnaissance du type de véhicules (LPR Editor). L'objectif de ce projet est d'étudier différents algorithmes d'analyse d'images pour la détection de véhicules obstacles. Ces derniers sont situés principalement en secteur avant du véhicule porteur (du système de capteurs). La seconde partie de l'étude porte sur la classification du type de véhicule détecté dans une des classes suivantes : véhicules de tourisme, camionnettes, petits camions (3.5t), gros camions et semi-remorques (type 39t) et bus.

4.1.1 Contexte

L'augmentation du parc automobile s'accompagne d'une demande croissante de systèmes d'aide à la conduite, promettant une conduite plus sécurisée et confortable. Dans cette optique, de nombreuses recherches ont été menées par la communauté ITS (*Intelligent Transport Systems*). Ceci se traduit par l'installation de dispositifs de hautes technologies dans les véhicules et sur la route. Dans ce cadre, les systèmes de vision embarqués sont couramment utilisés. Ils peuvent en effet fournir une description de la localisation et de la taille des autres véhicules dans l'environnement, ainsi que de la route, des panneaux de signalisation et des autres usagers de la voirie. Une détection des véhicules obstacles en temps réel est exigée pour ce type d'applications, de manière à laisser au conducteur le temps de réagir. La recherche exhaustive des positions potentielles des véhicules dans l'image complète est prohibitive pour les applications en temps réel. Pour résoudre ce problème, la plupart des méthodes de la littérature se décomposent en deux étapes, selon un processus attentionnel :

- Génération des Hypothèses : le système fournit, à l'aide d'algorithmes de traitement d'images simples et rapides, des positions potentielles de véhicules afin de restreindre le champ de recherche.
- Validation des Hypothèses : les hypothèses issues de l'étape antérieure sont validées en utilisant des algorithmes de reconnaissance des formes plus complexes et les fausses détections sont éliminées.

La génération d'hypothèses se base sur des informations comme la texture, la symétrie horizontale, la couleur, l'ombre portée, les segments, la route … D'autres méthodes, basées sur la stéréo ou le mouvement ont été proposées, qui n'entrent pas dans le cadre de cette étude. Les approches mises en œuvre pour valider les hypothèses s'appuient soit sur des primitives (gabarits rigides ou déformables), soit sur l'apparence. Dans ce cas, les caractéristiques de la classe véhicule sont apprises sur une base d'images. Chaque image est représentée par un vecteur de caractéristiques locales ou globales (contours orientés, ondelettes de Gabor, de Haar …). Puis, l'apprentissage d'un classifieur permet d'estimer la frontière de décision entre la classe « véhicule » et la classe « non-véhicule ». Un état de l'art récent [SBM06[105]] présente en détail toutes ces méthodes.

[105] Sun Z., Bebis G. & Miller R., On-road vehicle detection: A review. IEEE Transactions on Pattern Analysis and Machine Intelligence, 28(5) :694–711, 2006.

4.1.2 Travaux réalisés [Annexe 6]

Nous avons choisi d'approfondir les techniques de boosting et d'utiliser une cascade attentionnelle similaire à celle proposée par Viola & Jones. Toutefois, au lieu de mettre en œuvre un seul jeu de caractéristiques, nous en avons combiné plusieurs. L'idée n'est pas nouvelle et des travaux contemporains proposent aussi la fusion de plusieurs descripteurs : aux ondelettes de Haar sont associés des ondelettes de Gabor [SBM05[106]] ou des histogrammes de gradient orienté [GLP07[107]]. Mais les classifieurs faibles associés à ces caractéristiques, sont tous discriminants. L'originalité de notre approche repose sur le fait d'avoir combiner des classifieurs faibles discriminants et génératifs. Nous en verrons tout l'intérêt par la suite.

La base de données, fournie par PSA, contient plus de 800 images contenant une ou plusieurs vues arrières de véhicules, pour un total de 1200 images de véhicules de tourisme et utilitaires. Elle a été étiquetée par nos soins et chaque véhicule est repéré par un rectangle englobant. De cet ensemble ont été extraites les bases (apprentissage, validation croisée et test) d'exemples positifs. La base d'exemples négatifs est constituée d'images tirées aléatoirement dans un ensemble de 4000 images quelconques (ne contenant pas de véhicules).

4.1.2.1 Caractéristiques utilisées

Nous comparons deux types de caractéristiques avant de les combiner : les filtres rectangulaires et les histogrammes de gradient orienté. Ces deux descripteurs sont actuellement utilisés couramment par la communauté reconnaissance de formes pour la détection ou la reconnaissance d'objets.

Filtres de Haar

Les premiers descripteurs ont été initialement introduits dans [PP00[108]] pour la détection de véhicules en s'inspirant d'une décomposition via des ondelettes de Haar. Cette base de filtres a été élargie par la suite dans différents travaux [LKP03[109]], qui ne suivent plus strictement la théorie des ondelettes. Ils sont alors nommés filtres rectangulaires ou filtres de Haar (*Haar-like filters*) et donnent une information sur la distribution des niveaux de gris dans deux régions voisines de l'image. Une représentation intermédiaire de l'image, appelée image intégrale, permet d'évaluer ces filtres très rapidement. Comme on peut le constater (Figure 28.a), ces filtres mettent en valeur les contours (verticaux ou horizontaux) de l'image. Le changement d'échelle se traduit par un filtrage des détails qui ne conserve que les principaux contours. L'espace des paramètres de Haar est donc défini par un vecteur contenant 8151 caractéristiques pour une vignette de taille 32x32 pixels. Chaque caractéristique j est définie par : $f_j = (x_j, y_j, s_j, r_j)$, où r_j est le type de filtre rectangulaire, s_j l'échelle (nous en considérons 5 au total) et (x_j , y_j) la position. Nous définissons la fonction de classification faible associée au descripteur j du vecteur de paramètres, comme une réponse binaire (g_{Haar} =1 si $p_j f_j < p_j \theta_j$; 0 sinon) où f_j est la valeur du descripteur, θ_j est le seuil et p_j est la parité. Pour chaque descripteur, l'apprentissage détermine le seuil θ_j optimal qui minimise le nombre d'exemples mal classés (non détections et fausses alarmes).

---

[106] Sun Z., Bebis G. & Miller R., On-road vehicle detection using evolutionary gabor filter optimization, IEEE Transactions on Intelligent Transportation Systems, 2005.

[107] Geronimo D., Lopez A., Ponsa D. & Sappa A.D., Haar wavelets and edge orientation histograms for on-board pedestrian detection, IbPRIA, pp 418–425, 2007.

[108] Papageorgiou C. & Poggio T., A trainable system for object detection, International Journal of Computer Vision, 38(1): 15–33, 2000.

[109] Lienhart R., Kuranov A. & Pisarevsky V., Empirical analysis of detection cascades of boosted classifiers for rapid object detection, DAGM03, pp 297–304, 2003.

Histogrammes de gradient orienté

Les HoG appartiennent au deuxième groupe de caractéristiques utilisées dans ce travail. Dans des travaux récents [DT05[110]], organisés sous la forme de descripteurs SIFT (*Scale Invariant Feature Transform*) [L99[111]], ils ont été utilisés avec succès pour la détection de piétons. Ils utilisent le module et l'orientation du gradient autour d'un point d'intérêt ou à l'intérieur d'une région de l'image pour construire un histogramme. Nous observons (Figure 28.b), que dans une des régions, les points de contours trouvés sont, en majorité, d'orientation horizontale (classe deux de l'histogramme). L'autre région, contient des pixels de contours de toutes les classes, avec toutefois beaucoup de points verticaux. L'espace de paramètres de HoG est donc défini par un ensemble contenant 3917 histogrammes définis, comme précédemment, par $h_j = (x_j, y_j, s_j, r_j)$, pour une vignette de 32x32 pixels. Cette fois, nous estimons la distance entre un histogramme h_j de l'image d'entrée et un histogramme modèle m_j. Ce modèle est la valeur médiane des histogrammes de tous les exemples d'apprentissage positifs. Nous définissons alors le classifieur faible ($g_{HoG} = 1$ si $d(h_j, m_j) < \theta_j$; 0 sinon) où d est la distance de Bhattacharya entre l'histogramme d'entré h_j et l'histogramme du modèle m_j, et θ_j est le seuil optimal sur la distance pour cette caractéristique.

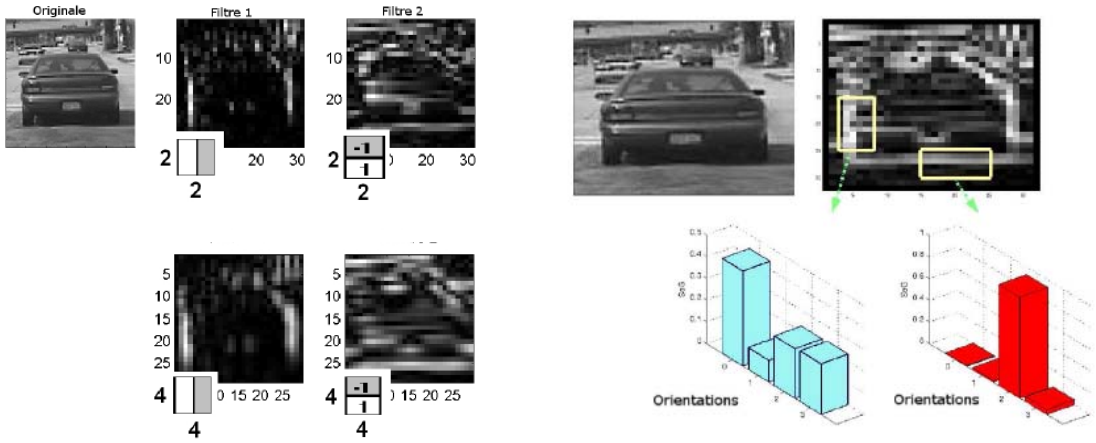


Figure 28. Application des filtres de Haar et des HoG à l'image d'un véhicule.

[110] Dalal N. & Triggs B., Histograms of oriented gradients for human detection, International Conference on Computer Vision and Pattern Recognition, (2), pp 886–893, 2005.

[111] Lowe D.G., Object recognition from local scale invariant features, International Conference on Computer Vision, pp 1150–1157, 1999.

4.1.2.2 Schémas de détection

Nous avons mis en œuvre et comparons plusieurs détecteurs de complexité croissante. Au niveau des caractéristiques d'abord, nous évaluons les filtres de Haar, les histogrammes de gradient orienté et leur concaténation. Au niveau de l'architecture ensuite, nous utilisons successivement un détecteur simple constitué d'un seul classifieur fort, puis, une cascade attentionnelle de classifieurs forts, enfin, une cascade « contrôlée », où nous limitons le nombre de caractéristiques. Les indices de performances considérés sont les suivants. Le taux de détections correctes *DC* correspond au pourcentage de véhicules correctement détectés sur l'ensemble des véhicules présents dans les images de la base de test. Le taux de fausses alarmes *FA* est calculé à partir de la moyenne des fausses alarmes par image (calculée sur l'ensemble des images de la base de test) divisée par le nombre total de vignettes évaluées par le détecteur dans une image (31514). Enfin, le temps de traitement moyen par image est évalué sur un PC cadencé à 2.2GHz.

Détecteur simple

Il est composé d'un classifieur fort *G* unique, construit avec un nombre *T* de classifieurs faibles. L'apprentissage a pour objectif d'atteindre un taux de détection donné sur la base de validation (99.5%). Les performances du détecteur sont résumées dans le Tableau 12.

Type	#Desc	DC(%)	FA	Temps(sec)
Haar	50	99.8	0.0220	1.42
Haar	100	99.8	0.0145	3.51
Haar	150	99.0	0.0044	5.17
HoG	50	100	0.0588	0.90
HoG	100	99.9	0.0300	1.68
HoG	150	99.9	0.0233	2.33
Fusion	50	99.6	0.0130	1.67
Fusion	100	99.3	0.0093	3.19
Fusion	150	99.2	0.0063	4.75

Tableau 12. Performances du détecteur simple.

Le fait d'augmenter le nombre de caractéristiques (*#Desc*) raffine la frontière, dans le cas des filtres de Haar, et le modèle, dans le cas des HoG. Du point de vue des fausses alarmes, les descripteurs de Haar se montrent plus discriminants que les descripteurs de HoG. La fusion a un comportement intermédiaire, en conservant un taux de détections élevé tout en éliminant un grand nombre de fausses alarmes. Nous observons aussi que le temps de calcul par image augmente naturellement avec la quantité de descripteurs. L'utilisation d'un grand nombre de descripteurs devient irréaliste du point de vue d'une application en temps réel. D'où l'intérêt de l'architecture en cascade.

Cascade attentionnelle

Elle est composée d'une succession des classifieurs forts G_i. Chaque classifieur fort de la cascade est entraîné en utilisant l'algorithme Adaboost. Au lieu de borner la boucle itérative d'AdaBoost par le nombre maximum de caractéristiques (*T*), nous fixons deux objectifs : le taux minimum de détections correctes DC_{min} (99.5%) et le taux maximal de fausses alarmes acceptables FA_{max} (0.4). La base négative N_i, utilisée pour entraîner le classifieur fort G_i de l'étage *i*, est formée d'exemples négatifs ayant été mal classés (c'est-à-dire considérés comme des véhicules) par les étages précédents.

Type	#Nég	#étages	#Desc	DC(%)	FA	Temps	Arrêt
Haar	1000	12	430	95.4	0.00080	0.59	Non conv
Haar	2000	11	479	96.4	0.00070	0.57	Non conv
Haar	3000	10	272	97.7	0.00099	0.58	Non conv
HoG	1000	5	89	99.8	0,030	0.73	Non conv
HoG	2000	5	52	99.9	0,034	0.56	Non conv
HoG	3000	4	21	99.9	0,077	0.43	Non conv
Fusion	1000	14	392	94.5	0.00027	0.39	F atteint
Fusion	2000	12	369	93.9	0.00035	0.37	Non conv
Fusion	3000	12	358	94.3	0.00039	0.36	Nég. non atteint

Tableau 13. Performances de la cascade.

Nous observons tout d'abord une disparité importante au niveau du nombre d'étages de la cascade entre les trois détecteurs. La plupart des apprentissages se sont arrêtés du fait d'une non-convergence de l'algorithme AdaBoost. Lorsque nous augmentons le nombre de vignettes négatives, les processus ne convergent plus pour un nombre d'étages plus faible. Ceci s'explique aisément : une base négative de grande taille permet de générer un modèle ou une frontière suffisamment robuste pour éliminer dès les premiers étages de la cascade un grand nombre de fausses alarmes. Très vite, il ne reste que des cas difficiles à rejeter, d'où la non convergence de l'algorithme. Remarquons aussi que les détecteurs basés sur les HoG convergent avec un nombre nettement plus faible de caractéristiques que ceux utilisant les Haar. En effet, il faut peu de caractéristiques pour éliminer les vignettes négatives relativement éloignées du modèle. Inversement, le détecteur de Haar nécessite un plus grand nombre de caractéristiques pour estimer correctement la frontière entre les classes. A nouveau, nous pouvons constater que le détecteur de fusion permet d'améliorer les performances. Nous vérifions enfin que le nombre de fausses alarmes est fortement lié au nombre d'étages dans la cascade. Une quantité plus importante d'étages permet d'éliminer plus de fausses alarmes.

Cascade contrôlée

Afin d'éviter une non-convergence de la cascade, nous ajoutons un critère d'arrêt portant sur le nombre maximum de caractéristiques dans l'apprentissage de la fonction G_i. Pour fixer cette grandeur à chaque étage de la cascade, nous pouvons utiliser une loi croissante à notre convenance (exponentielle, par exemple). Le tableau 3 montre les performances globales des détecteurs en cascade contrôlée. Le détecteur de HoG obtient un taux de détections correctes important, cependant le taux de fausses alarmes reste aussi élevé. Celui de Haar a un comportement inverse. A la fin de la cascade, il obtient une faible quantité de fausses alarmes, mais de nombreux exemples positifs ont été éliminés dans les étages précédents. Le détecteur Fusion combine les avantages des caractéristiques génératives de HoG et discriminantes de Haar. Dans les premiers étages (Figure 29), il utilise essentiellement les caractéristiques génératives pour éliminer les échantillons négatifs éloignés du modèle en conservant un taux de détections élevé ; puis, dans les derniers étages, il se sert des caractéristiques de Haar pour estimer des frontières nettes entre les exemples positifs et les négatifs encore présents (les plus proches de la frontière). Ce comportement se traduit par le rejet d'un grand nombre de fausses alarmes dès les premiers étages de la cascade et une diminution significative du temps de réponse.

Type	# Neg	# Desc	DC (%)	FA	t (seg)
Haar	1000	1016	93.8	0.00031	0.66
Haar	3000	942	89.83	0.00018	0.69
HoG	1000	1027	97.8	0.0045	0.51
HoG	3000	1031	99.6	0.0114	1.07
Fusion	1000	1022	94.0	0.00029	0.36
Fusion	3000	1021	93.5	0.00032	0.40

Tableau 14. Performances de la cascade contrôlée.

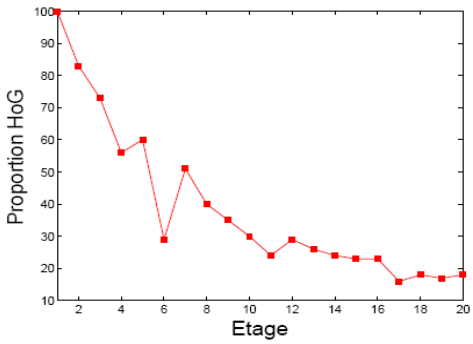


Figure 29. Proportion des caractéristiques de HoG utilisée dans le détecteur fusion à cascade contrôlée.

Ceci se reflète très concrètement dans les images de scènes routières de la Figure 30, où les carrés blancs signalent une détection (correcte ou fausse) obtenue. Nous observons que le détecteur de Haar ne détecte pas tous les véhicules mais ne produit que peu de fausses alarmes. En revanche, le détecteur de HoG réalise de nombreuses fausses alarmes, mais détecte quasiment tous les véhicules. Enfin, le détecteur de fusion réduit sensiblement le nombre de fausses alarmes tout en détectant les mêmes véhicules que celui de HoG.

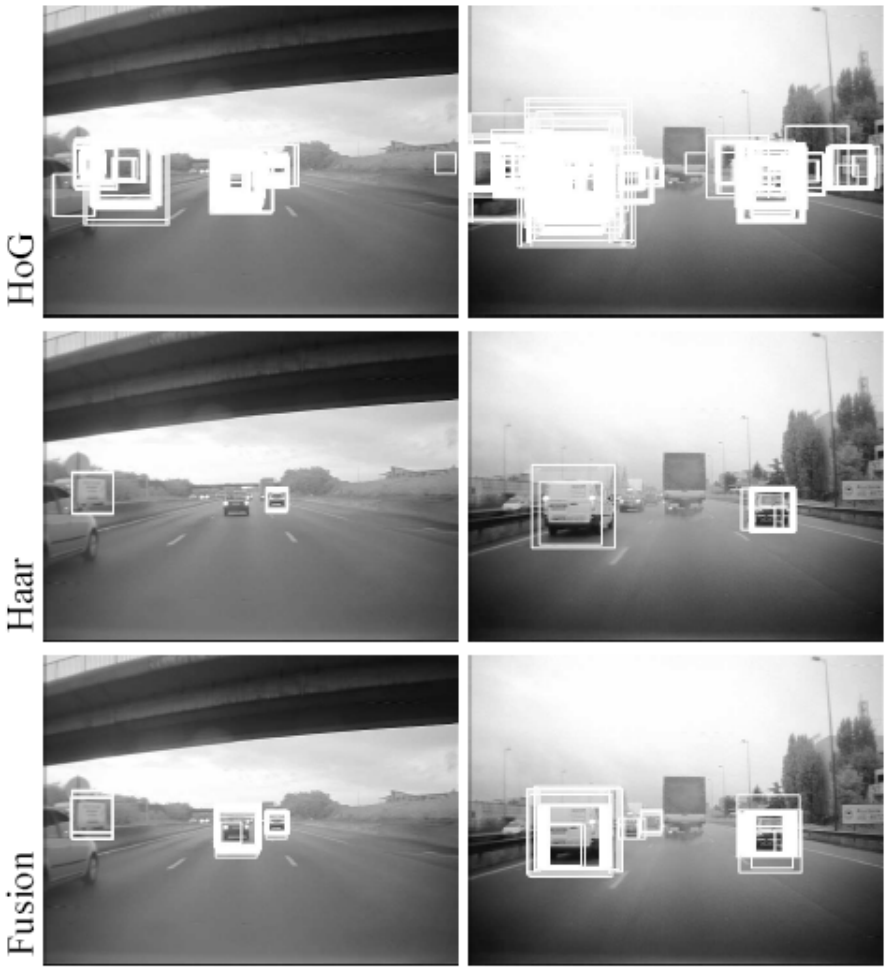


Figure 30. Exemples du comportement des trois détecteurs sur des images de scènes autoroutières.

4.1.3 Conclusions

Dans cette étude, nous avons proposé deux espaces de paramètres, les filtres de Haar et les histogrammes de gradient orienté (HoG), appliqués à la détection de véhicules, en utilisant une approche similaire à celle de Viola-Jones. Les premiers sont associés à des classifieurs faibles de type discriminants et les seconds à des classifieurs faibles de type génératifs. Un troisième détecteur est obtenu à partir d'une fusion de ces deux jeux de caractéristiques. Nous avons étudié le comportement de différentes architectures : le détecteur simple et le détecteur en cascade. Pour optimiser la performance de ce dernier, nous avons fixé le nombre maximum de caractéristique dans chaque étage de la cascade. Le détecteur réalisant la fusion des deux caractéristiques combine les avantages des précédents détecteurs (de Haar et de HoG) : un taux de détections correctes élevé et un faible nombre de fausses alarmes. Il utilise les classifieurs de type génératifs pour éliminer les échantillons négatifs éloignés du modèle, puis il se sert des descripteurs discriminants pour dessiner des frontières nettes entres les exemples positifs et ceux négatifs proches, encore présents. Nos travaux futurs seront consacrés à la classification du type de véhicules. Quelques pistes de recherches dans ce domaine seront proposées au paragraphe §5.1.

4.2 LOCALISATION DE TEXTES

Encadrements : 1 thèse (S. Muhammad Hanif), 1 stage (L. Nguepdjop)

Ces recherches ont commencé fin 2006 et s'insèrent dans les thématiques du nouvel Institut des Systèmes Intelligents et Robotiques (ISIR, CNRS FRE2507), créé le 1er janvier 2007. Ce travail fait partie du projet « lunettes intelligentes » dont le but est d'aider les aveugles et les personnes mal voyantes à mieux connaître leur environnement. « Lunettes intelligentes » (Figure 31) est un système associant un banc de stéréovision, une centrale inertielle (constituant avec un système de traitement approprié, un système de perception visuelle) et une surface à stimulation tactile de la main de type surface Braille [RMP03[112]]. Ce système offrira aux personnes aveugles ou mal-voyantes une représentation tactile de leur environnement. La surface tactile du système est une représentation de l'espace établie à partir de l'information visuelle (acquise par le système de stéréovision) et pondérée par l'information inertielle. Elle varie avec le temps et le point d'observation de la scène par le sujet. Il est également possible d'estimer les relations mesurables de l'espace (comme la taille d'un objet, la distance à un obstacle, ….). Le besoin d'information textuelle est évident pour les aveugles et mal voyantes. Nous avons choisi d'ajouter un module supplémentaire à l'étape de perception du système, afin de détecter, localiser et reconnaître le texte dans les images capturées. Cette information textuelle particulièrement riche (nom de rues, enseigne de magasins …) sera représentée sur la surface tactile.

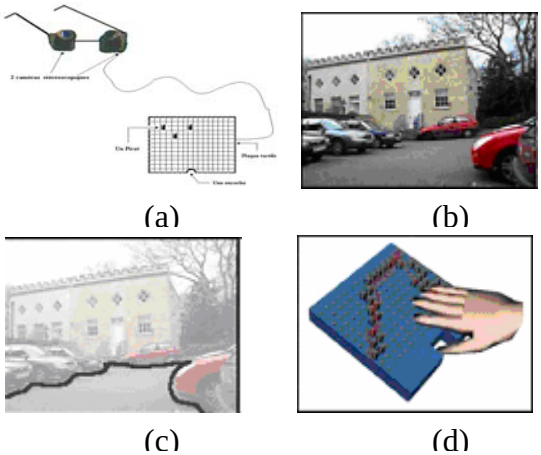


Figure 31. Lunettes Intelligentes: (a) conception (b) scène (c) perception d'environnement (d) représentation tactile associée.

4.2.1 Contexte

Les problèmes de détection et de reconnaissance de textes dans les images « naturelles » ont été abordés récemment par les chercheurs de la communauté DAR. Ainsi, le premier workshop portant sur le sujet (*Camera-Based Document Analysis and Recognition* – satellite d'ICDAR) a eu lieu en 2005. Les avantages et inconvénients des capteurs d'images sur les traditionnels scanners sont résumés dans [LDL03[113]] : un très grand nombre d'utilisateurs potentiels et un usage facilité (en raison de l'intégration du capteur dans les téléphones

[112] Velázquez R., Maingreaud F. & Pissaloux E., Intelligent Glasses: A New Man-Machine Interface Concept Integrating Computer Vision and Human Tactile Perception, EuroHaptics 2003, Dublin, Ireland, July 2003.
[113] Liang J., Doermann D. & Li H., Camera-based analysis of text and documents: a survey, International Journal on Document Analysis and Recognition, 7(2-3), pp 84-104, 2005.

portables, PDAs et autres appareils « nomades ») d'une part ; un ensemble de problèmes nouveaux à résoudre (basse résolution, distorsion perspective, flou, fond complexe …) d'autre part. La plupart des algorithmes présentés dans la littérature suivent le paradigme présenté dans la section précédente. Ils commencent par une phase de génération d'hypothèses (détection) où sont estimées les régions de l'image susceptibles de contenir du texte. Cette phase utilise des méthodes de traitement d'images qui peuvent grossièrement être classifiées en trois catégories principales s'appuyant sur le gradient, la couleur ou la texture. Suit une phase de validation d'hypothèses où les fausses alarmes sont éliminées et les zones de texte localisées dans l'image. Les méthodes les plus couramment utilisées à ce stade effectuent un étiquetage en composantes connexes de l'image de détection suivi d'une analyse des caractéristiques géométriques des composantes (taille, ratio, alignement …). Finalement, des algorithmes d'amélioration d'images (correction perspective, entre autres) et de reconnaissance permettent d'identifier le texte.

4.2.2 Travaux réalisés

Nous nous sommes pour l'instant concentrés sur l'étape de génération des caractéristiques, plus particulièrement sur les méthodes de détection basées sur la texture. En effet, le texte possède une texture unique qui montre une certaine régularité, facile à distinguer du fond. Les êtres humains peuvent ainsi identifier un texte écrit en langue étrangère même s'ils ne comprennent pas cette dernière, et ce grâce à sa texture distincte. Des chercheurs ont exploité ce fait pour détecter le texte dans les images. Les caractéristiques de texture peuvent être directement extraites à partir des données brutes (pixels). Nous avons dans un premier temps utilisé les matrices de co-occurrence des niveaux de gris (MCNG) et, comme [HSD73[114]], avons extrait de ces dernières plusieurs paramètres (contraste, homogénéité, dissimilarité, entropie, énergie et corrélation). Puis, sachant que la matrice de co-occurrence représentait la probabilité conjointe $p(i,j)$ de deux pixels, nous avons considéré les distributions marginales $p(i)$ ou $p(j)$ liées à cette probabilité conjointe. La MCNG étant calculée pour différentes orientations et différentes distances, elle couvre les relations spatiales entre pixels. Les distributions marginales (appelées par la suite histogramme spatiaux) doivent donc contenir à la fois des informations de texture, de forme et des relations spatiales. Elles peuvent être utilisées pour la classification. Nous avons évalué les propriétés des deux représentations de formes (paramètres de texture et histogrammes spatiaux) à l'aide de plusieurs classifieurs bayésiens (avec modèles mono ou multi-gaussien de la classe « texte » seule ou des classes « texte » et « non-texte ») et neuronaux.

Nous avons utilisé une partie de la base d'images ICDAR 2003[115] (couramment utilisée dans les compétitions de localisation de textes) pour les expérimentations. Ces images contiennent des textes de diverses tailles de police, différentes longueurs de mots, orientations et couleurs. Nous avons constaté que modéliser les deux classes donne de meilleurs résultats que modéliser uniquement la classe texte. De plus, le modèle uni-gaussien fonctionne mieux que le modèle multi-gaussien. Enfin, le classifieur neuronal est le plus performant. En comparant les deux techniques de codage de texture, nous constatons que le détecteur basé sur l'histogramme spatial est meilleur – augmentation de 2% en taux de détection (qui atteint 66%) et diminution de 3% du taux de fausses alarmes (22%). De plus, la technique de l'histogramme spatial est plus rapide : le temps de calcul moyen des paramètres de MCNG sur une image de taille 480x640 est 196 secondes alors qu'il n'est que de 135 secondes pour l'histogramme spatial. La Figure 32 présente quelques résultats de détection.

[114] Haralick R.M., Shanmugam K. & Dinstein I., Textual Features for Image Classification, IEEE Transactions on Systems, Man, and Cybernetics, 3(6), pp.610-621, 1973.
[115] ICDAR 2003 Robust Reading and Text Locating Competition,http://algoval.essex.ac.uk/icdar/RobustReading.html.

Figure 32. Exemples de détection de texte par la technique des histogrammes spatiaux.

4.2.2 Conclusions

Nous avons proposé un codage simple de texture pour la détection de texte dans des images de scène naturelle. Nous avons constaté que les histogrammes spatiaux estimés à partir de la matrice de co-occurrence sont plus efficaces que les paramètres de texture pour la détection de texte. Bien que la performance soit évaluée sur une petite base (50 images de test), les résultats sont encourageants et nous espérons que l'évaluation de performance sur une grande base de test validera ces résultats et conclusions. Nous somme en train d'explorer d'autres méthodes de détection de texte basées sur le calcul du gradient de l'image. Ces méthodes sont en effet réputées plus rapide que les précédentes. Nous travaillons aussi sur l'étape de validation, afin d'éliminer les zones de « non-texte » détectées.

5. PERSPECTIVES

5.1 RETOUR SUR LE PASSE POUR MIEUX PREPARER L'AVENIR …

D'un point de vue méthodologique, comme le lecteur a pu le constater, nos contributions se situent essentiellement dans le domaine de la **fusion d'informations**. Elles se déclinent toutefois selon plusieurs stratégies que nous rappelons dans la suite, en omettant cette fois l'ordre chronologique de nos recherches, afin de mettre en valeur leur cohérence.

Plusieurs méthodes de **coopération guidée par les données** ont été étudiées. La première a été proposée dans le cadre de la reconnaissance d'écriture, plus précisément la classification de caractères isolés. Nous avons constaté (§3.1.3) les limites des classifieurs dédiés respectivement à l'analyse du tracé du caractère (représentation dynamique) et de l'image du caractère (représentation statique) dans le cadre omni-scripteur, particulièrement complexe en raison de la grande variabilité et de l'hétérogénéité du signal à reconnaître. Si les performances des classifieurs associés à ces deux représentations n'étaient pas suffisantes, leur comportement était satisfaisant. En effet, si l'un d'eux échouait à classer correctement une donnée, l'autre réussissait souvent. Cette « orthogonalité » permettait d'espérer que la coopération des deux classifieurs entraînerait une amélioration globale des performances. Cette hypothèse a été doublement confirmée. D'un point de vue théorique d'abord, car plusieurs analyses approfondies des systèmes de combinaison rapportées dans la littérature ont montré sa validité. D'un point de vue expérimental ensuite, car toutes les combinaisons que nous avons mises en œuvre ont été couronnées de succès. Elles ont permis de concevoir un système de reconnaissance de caractères isolés particulièrement performant. La même démarche a prévalu en localisation de visages (§3.3.2). Nous avons dans un premier temps développé un détecteur basé sur les seules propriétés anthropomorphiques du visage. Mais la variabilité de ces dernières, due à la multitude des orientations et des expressions faciales, limitait la robustesse du détecteur. Nous avons alors choisi de suivre le même paradigme « diviser pour mieux régner » et conçu un système modulaire recherchant dans l'image des informations géométrique (forme elliptique du visage) et colorimétrique (teinte « chair »). Une fois de plus, les conflits étaient nombreux entre ces trois sources d'informations. Leur combinaison a permis l'émergence d'une décision finale plus pertinente. Dans les deux cas, nous avons extrait du signal (mono ou bidimensionnel) à traiter des représentations différentes suffisamment décorrélées pour être combinées.

Nos travaux en localisation des caractéristiques faciales (§3.3.3) suivaient apparemment le même paradigme. Constatant les limites d'un unique expert pour traiter le problème dans son ensemble (quelque soit l'orientation du visage), nous avons décomposé ce dernier en sous-problèmes plus simples à résoudre. Ainsi, nous avons entraîné plusieurs experts, spécialisés sur un ensemble donné d'orientations. Puis, nous avons mis en compétition les hypothèses émises par ces experts afin de sélectionner la plus cohérente. Le même paradigme a donc été utilisé. Toutefois, au lieu de s'appuyer sur des représentations différentes liées aux données traitées (autrement dit, de « projeter » les données initiales dans différents espaces de représentation), nous avons séparé l'espace des caractéristiques initiales de façon non supervisée avant d'entraîner un expert sur chaque domaine défini par la partition. Cette

solution, satisfaisante car indépendante des données, peut s'appliquer dans un cadre plus général. Elle n'est pas sans rappeler les *k*-plus-proches-classifieurs proposés dans [PE04[116]].

De la même façon, les travaux menés dans le cadre de la reconnaissance d'écriture (§3.1.4), préconisant une **combinaison hiérarchique de classifieurs génératif et discriminant** sont génériques et applicables à n'importe quel problème de classification. Le premier niveau confie à l'étage génératif le soin d'éliminer les hypothèses les moins pertinentes, afin que le second étage, discriminant, se concentre sur les données les plus difficiles à séparer. Cette solution, *a priori* intuitive, a été validée expérimentalement en classification de caractères, mais aussi en détection de véhicule. Nos dernières expériences ont en effet montré que, dans un processus de classification hiérarchique, les caractéristiques génératives étaient utilisées en premier, suivies par les caractéristiques discriminantes.

Les précédentes contributions combinaient, à l'aide d'architectures parallèle ou séquentielle, des informations homogènes, typiquement des probabilités d'appartenance à une classe (caractère, teinte « chair » …). Les travaux que nous avons conduits en lecture automatique de textes manuscrits (§3.2.2) s'appuyaient au contraire sur la **fusion d'informations hétérogènes**. Ainsi, plusieurs niveaux d'abstraction étaient analysés. Des informations « bas niveau » géométriques (silhouette des caractères) ou topologiques (agencement des caractères/mots les uns par rapport aux autres), coexistaient avec des informations de plus « haut niveau », lexicales ou phonétiques (dans le cas du modèle à triple voie). Nous avons montré que le cadre statistique était parfaitement adapté pour combiner ces différentes sources.

Enfin, récemment, les **techniques de boosting** nous ont offert un cadre intéressant pour la résolution de problèmes de détection d'objets (§4.1). L'algorithme initial (*discrete Adaboost*) proposait une combinaison par somme pondérée des réponses de T classifieurs faibles, pour créer un classifieur fort G. Sa mise en œuvre a montré que les performances, sur un problème de détection, s'amélioraient lorsque T croissait, au prix d'une augmentation significative du temps de traitement, linéaire en T. Afin d'éliminer cette imperfection, nous avons choisi de construire itérativement une cascade attentionnelle de classifieurs forts G_i et d'analyser en détail son comportement. Nos conclusions rejoignent celle des concepteurs de l'algorithme : la cascade permet de diminuer sensiblement le taux de fausses alarmes et le nombre de caractéristiques utilisées (donc le temps de traitement), au prix d'une légère diminution du taux de détection. Associée à la cascade, l'utilisation de plusieurs jeux de caractéristiques a montré tout son intérêt. En particulier, nous avons vérifié expérimentalement que certaines caractéristiques étaient utilisées dans les premiers étages de la cascade, afin d'éliminer le plus grand nombre de fausses alarmes, alors que d'autres n'entraient en jeu que dans les derniers étages, en vue de raffiner la frontière de décision en s'appuyant sur les exemples les plus proches de celle-ci (on retrouve ici les vecteurs supports des machines du même nom et la notion de maximisation de la marge).

[116] Prudent Y. Ennaji A., La topologie des données pour une distribution fiable des tâches de classification, 11èmes Rencontres de la Société Francophone de Classification, pp 282-285, 2004.

Le **développement de techniques dérivées du boosting** (par abus de langage, nous inclurons dorénavant dans cette expression la cascade attentionnelle, afin d'éviter l'usage de l'élégant acronyme CoBE pour *Cascade of Boosting Ensembles*) mérite d'être poursuivi pour de nombreuses raisons. La multiplication de travaux récents sur le sujet rapportés dans la littérature nous y incite. Plusieurs extensions ont retenu notre attention :

- L'algorithme *FloatBoost* [LZ04[117]] préconise une **modification de la phase d'extraction des caractéristiques**. Plus précisément, il applique une sélection par *Floating Search*. L'algorithme initial enchaîne les phases *Forward* (ajout des classifieurs faibles les plus pertinents) jusqu'à satisfaction de l'objectif. Au contraire, celui-ci alterne des phases *Forward* et *Backward* où les classifieurs faibles à l'origine d'une baisse de performance sont supprimés. Les expériences montrent qu'à nombre de classifieurs faibles constant, les performances de l'algorithme *FloatBoost* sont meilleures que celle d'*AdaBoost* ; autrement dit, il réclame moins de classifieurs faibles pour atteindre des performances similaires. D'autres méthodes de filtrage sont proposées dans [BMR06[118]] (basées sur la mesure du pouvoir discriminant ou la maximisation de l'information mutuelle des caractéristiques) qui, utilisées en amont de l'apprentissage proprement dit, atteignent les mêmes objectifs.

- [HAW04[119]] propose quant à lui une **modification de l'architecture de la cascade attentionnelle**. Le premier classifieur faible utilisé pour construire le classifieur fort G_i est la sortie du classifieur fort précédent G_{i-1}. Là encore, les expériences rapportent une amélioration significative des performances : meilleur compromis taux de détection/fausses alarmes et nombre plus faible de classifieurs faibles utilisés quelque soit l'étage de la cascade synonyme d'un traitement plus rapide.

- Le **choix de la fonction de classification faible** est aussi une problématique intéressante. Le succès initial des filtres de Haar était principalement dû à la possibilité d'évaluer très rapidement la sortie d'un grand nombre de filtres en ayant recours à l'image intégrale. De même, l'utilisation des histogrammes de gradient orienté a été facilitée par l'histogramme intégral. Les fonctions de classifications faibles associées à ces caractéristiques étaient des plus simples : binaires et discrètes dans l'algorithme initial, puis continues pour *Real AdaBoost*, elles n'indiquaient que la (probabilité de) présence d'une caractéristique pertinente pour la tâche à réaliser. Nous avons proposé de panacher des classifieurs génératifs et discriminants .D'autres mettent en œuvre des classifieurs « moins » faibles : arbres de décision C4.5 [FS96[120]] ou CART [BMR06] SVM [ZYC06[121]],…. Une version récente (*MBoost*) [ZI07[122]] propose de sélectionner les classifieurs faibles dans un ensemble comprenant des classifieurs bayésiens, des classifieurs directs, des arbres décisions et des machines à vecteurs supports.

---

[117] Li S.Z. & Zhang Z., Floatboost learning and statistical face detection, IEEE Transactions on Pattern Analysis and Machine Intelligence, 26(9):1112-1123, 2004.

[118] Brubaker S.C., Mullin M.D. & Rehg J.M., Towards optimal training of cascaded detectors, European Conference on Computer Vision, (1), pp 325-337, 2006.

[119] Huang C., Ai H., Wu B. & Lao S., Boosting nested cascade detector for multi-view face detection, International Conference on Pattern Recognition, (2), pp 415-418, 2004.

[120] Freund Y. & Shapire R.E., Experiments with a new boosting algorithm, International Conference on Machine Learning, pp 148-156, 1996.

[121] Zhu Q., Yeh M.C., Cheng K.T. & Avidan S., Fast human detection using a cascade of histograms of oriented gradients, International Conference on Computer Vision and Pattern Recognition, pp 1491–1498, 2006.

[122] Pang Z. & Isbell C., Managing domain knowledge and multiple models with boosting, International Joint Conference on Artificial Intelligence, 2007.

- Dans le domaine de la **classification (multi-classes)**, plusieurs extensions des techniques de boosting (initialement dédiées aux problèmes bi-classes) ont été proposées. Les premières réduisaient le problème multi-classes à un grand nombre de problèmes bi-classes [ASS00[123]]. Plus récemment, [ZRZ05[124]] reste en contexte multi-classe mais modifie l'algorithme AdaBoost et introduit dans la pondération des exemples d'apprentissage un facteur de pénalisation fonction du nombre de classes à séparer. [HAL07[125]] décrit différentes stratégies dérivées du boosting mises en œuvre pour détecter avec succès des visages quelque soit leur orientation : détecteur spécifiques à une orientation fonctionnant en parallèle, détecteur pyramidal se spécialisant au fur et à mesure de la cascade, arbres de décision … Toutes sont (plus ou moins aisément transposables) à la classification.

Toutes ces variantes augmentent les capacités de l'algorithme à traiter un problème complexe, au prix d'une augmentation significative du temps d'apprentissage. On retrouve finalement, dans le contexte du boosting, un **dilemme classique en reconnaissance des formes** : faut-il se concentrer sur la phase d'extraction des caractéristiques et mettre en œuvre des classifieurs (relativement) simples ou choisir une stratégie diamétralement opposée et faire porter tout le poids de la décision sur l'étage de classification ? Comme on a pu le constater, le boosting peut constituer la réponse à plusieurs questions posées au début de ce mémoire (§2.2). Ainsi, l'algorithme *MBoost* décrit précédemment montre que le boosting peut constituer une solution au problème du choix des classifieurs de base dans un système à classifieurs multiples. De même, les travaux présentés dans [LC07[126]] préconisent de n'utiliser le boosting que pour l'extraction de caractéristiques, suivi d'un SVM pour la classification.

Indépendamment des extensions évoquées ci-dessus, le boosting pose d'importantes questions quant au réglage de ses paramètres d'apprentissage. Quelle influence ont les objectifs fixés, le ratio entre exemples positifs et négatifs … sur la convergence et les performances de l'algorithme ? Enfin, reste le **problème crucial du choix de l'espace de représentation initial**. On l'a dit, le boosting combine les phases d'extraction de caractéristiques, de classification et de fusion afin de faire émerger des solutions dont les capacités de généralisation ont été prouvées, théoriquement et expérimentalement … mais échoue si l'ensemble de caractéristiques initial n'est pas adapté. Reste donc à trouver des méthodes de génération de caractéristiques « optimales ». Et, sur ce point, des recherches récentes tentent déjà de proposer des réponses riches de promesses [SRP07[127]].

Pour conclure, toutes les problématiques que nous venons d'évoquer constituent autant d'axes de recherches méthodologiques prometteurs pour les années à venir. Elles offrent aussi une occasion unique pour réunir des communautés de recherches différentes (extraction d'informations dans les données, apprentissage automatique, classification, fusion d'informations pour ne citer que les aspects méthodologiques) qu'on aimerait voir plus souvent … fusionner.

---

[123] Allwin E., R.E. Schapire & Singer Y., Reducing multi-class to binary: a unified approach to binary classifiers, Journal of Machine Learning Research, (1), 113-141, 2000.
[124] Zhu J., Rosset S., Zhou H. & Hastie T., Multi-class AdaBoost, Technical Report, 2005.
[125] Huang C., Ai H., Li Y. & Lao S., high-performance rotation invariant multiview face recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence, 29(4): 671-686, 2007.
[126] Leyrit L. & Chateau T., Association de classifieurs pour la sélection de variables, GdR ISIS, Journée détection et reconnaissance d'objet dans les images, 2007.
[127] Serre T., Riesenhuber M. & Poggio T., Robust object recognition with cortex-like mechanism, IEEE Transactions on Pattern Analysis and Machine Intelligence, 29(3), 2007.

5.2 APPLICATIONS VISEES

D'un point de vue applicatif, mes activités de recherches vont se distribuer sur deux problématiques bien distinctes.

Nos contributions dans le domaine de l'**analyse de visages** doivent être resituées au sein du thème « visage », en particulier les recherches conduites par M. Milgram sur le suivi du regard, la lecture labiale et la synthèse frontale. Tous ces travaux mettent en œuvre des techniques statistiques avancées d'analyse de visage (modèles de forme et d'apparence actifs) nécessitant, en amont, des algorithmes de détection de visages et des caractéristiques faciales. Dans le cadre plus vaste de l'équipe Perception et Mouvement du nouvel Institut, une action fédératrice intégrant les travaux sur la caractérisation du geste menés par C. Achard et X. Clady et ceux portant sur l'analyse du signal de parole conduits par J.L. Zarader et M. Chetouani est sur le point de naître. Les travaux menés récemment en analyse de visages et notre expertise en fusion d'informations trouveront naturellement leur place au sein de ces recherches. Elles seront menées au sein d'une **action COST** (Coopération européenne dans le domaine de la recherche scientifique et technique) visant à l'**analyse multimodale des communications verbale et non verbale**. On le sait, l'interaction humaine utilise deux voies distinctes. Le canal verbal véhicule des messages possédant un contenu sémantique spécifique; le canal non-verbal, des messages liés au sentiment général et à l'état émotionnel des personnes dialoguant. Si le canal verbal a déjà fait l'objet de nombreuses recherches, le rôle du canal non-verbal est moins bien compris. Les expressions faciales, les gestes et le regard constituent les principales composantes d'un « discours émotionnel » (information non-verbale) qui mérite d'être analysé. L'objectif principal de l'action est de développer l'analyse acoustique, perceptuelle, et psychologique avancée des signaux de communication verbaux et non-verbaux provenant de l'interaction en vis à vis spontanée, afin de développer des algorithmes capables d'identifier les états émotionnels. Parmi les différentes tâches de l'action, nous nous intéresserons en particulier à l'analyse des données multimodales. C'est-à-dire l'étude de la synchronisation existant entre les caractéristiques acoustiques, le mouvement des mains, du regard et les expressions faciales, en vue d'identifier les corrélations existant entre ces différentes caractéristiques.

Dans le domaine de l'**assistance à la personne**, les recherches que nous avons initiées récemment dans le cadre du projet « lunettes intelligentes » offrent de multiples perspectives. D'un point de vue méthodologique, le boosting a un fort potentiel pour résoudre les problèmes de détection de textes dans les images. L'idée initiale de **détection et reconnaissance de textes** pour l'aide à la navigation des déficients visuels peut déjà se décliner de diverses manières. A court terme, des scénarios « *indoor* » (d'intérieur) sont envisageables : très contraints comme le déplacement dans un laboratoire, un hôpital ou tout bâtiment administratif, où l'information pertinente se résume à des numéros de bureaux et des noms de services ou de personnes ; plus difficile, comme le déplacement dans les métro, gares et aéroports, où l'information comprend toute la « signalisation » : correspondances, plan de lignes .... Dans les deux cas, le problème est naturellement contraint de multiples manières : un éclairage (relativement) contrôlé limitant les risques de dégradation de l'image ; une faible variabilité des typographies rencontrées et un lexique dont le nombre d'entrées peut être grand mais reste dénombrable. Le scénario « *outdoor* » en environnement non contraint est beaucoup plus complexe : éclairage variable, multiplication des polices fantaisistes et lexique ouvert. De plus, nous venons de parler d'« information pertinente », se pose alors une question : quels critères utiliser pour décider de la « pertinence » d'une information ? C'est pourquoi d'autres applications moins ambitieuses pourront être envisagées : si l'on peut aider les déficients visuels à se déplacer, pourquoi ne pas leur permettre de « lire ». Toutes les

techniques évoquées plus haut, couplées à un capteur d'images et à l'interface tactile, permettent d'offrir, à faible coût, une représentation braille de n'importe quel livre. A plus longue échéance et dans une optique plus pédagogique, pourquoi ne pas envisager la transcription tactile de cours vidéo-projetés voire … de cours manuscrits.

Bref, de **nombreuses recherches fructueuses sur les plans méthodologiques, applicatifs et humains,** s'offrent à moi.

6. BIBLIOGRAPHIE

[A94] Akaike H., A new look at statistical model identification, IEEE Transactions on Automatic Control, 19: 716-723, 1994.

[ASS00] Allwin E., R.E. Schapire & Singer Y., Reducing multi-class to binary: a unified approach to binary classifiers, Journal of Machine Learning Research, (1), 113-141, 2000.

[AD91] Almuallim H. & Dietterich T.G., Learning with many irrelevant features, National Conference on Artificial Intelligence, pp 547-552, 1991.

[AL96] Anquetil E. & Lorette G. On-line Handwriting Recognition system Based on Hierarchical Qualitative Fuzzy Modeling, International Workshop on Frontier of Handwriting Recognition , pp 47-52, 1996.

[BC95] Baccino T. & Cole P., La lecture experte, Presse universitaire de France, 1995.

[BH89] Baldi P. & Hornik K., Neural networks and principal component analysis², Neural Network, 2(1):53-58, 1989.

[B95] Bishop, C. M., Neural Networks for Pattern Recognition. Oxford University Press, 1995.

[B03] Bloch I., Fusion d'informations en traitement du signal et des images, Lavoisier, 2003.

[BH 03] Bileschi S.M. & Heisele B., Advances in Component Based Face Detection, IEEE International Workshop on Analysis and Modeling of Face and Gestures, 2003.

[BM01] Brand J. & Mason J.S., Skin probability map and its use in face detection, International Conference on Image Processing, (1), 1034-1037, 2001.

[BMR06] Brubaker S.C., Mullin M.D. & Rehg J.M., Towards optimal training of cascaded detectors, European Conference on Computer Vision, (1), pp 325-337, 2006.

[B96] Breiman L., Bagging predictors, Machine Learning, 24(2), pp 123-140, 1996.

[CA04] Carbonnel S. & Anquetil E., Modélisation et intégration de connaissance lexicales pour le post-traitement de l'écriture manuscrite en-ligne, Congrès Reconnaissance des Formes et Intelligence Artificielle, Vol. 3, pp 1313–1322, 2004.

[CWS05] Chellappa R., Wilson C.L. & Sirohey S.: Human and machine recognition of faces: a survey, Proceedings of IEEE, 83(5), 705-740, 1995.

[C78] Coltheart M., Lexical access in simple reading task, Underwood, Strategies of information processing, Academic Press, 1978.

[C00] Connel S.D., On-line Handwriting Recognition Using Multiple Pattern Class Models, PhD thesis, Department of Computer science and engineering, Michigan State University, 2000.

[CJ02] Connell S. & Jain A.K. Writer adaptation of on-line handwriting models, IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(2): 329-342, 2002.

[CET98] Cootes T.F., Edwards G.J. & Taylor J.C., Active Appareance Models,. European Conference on Computer Vision, pp 484-498, 1998.

[CMK02] Cornujéols A., Miclet L. & Kodratoff Y., Apprentissage artificiel, Eyrolles, 2002.

[C97] Coté M., Utilisation d'un modèle d'accès lexical et de concepts perceptifs pour la reconnaissance d'images de mots cursifs, Thèse de doctorat, ENST, 1997.

[CC04] Cristinacce D., Cootes T.: A comparison of shape constrained facial feature detectors, International Conference on Automatic Face and Gesture Recognition, 375-380, 2004.

[DT05] Dalal N. & Triggs B., Histograms of oriented gradients for human detection, International Conference on Computer Vision and Pattern Recognition, (2), pp 886–893, 2005.

[DG05] Duffner S., Garcia C., A Connexionist Approach for Robust and Precise Facial Feature Detection in Complex Scenes, IEEE International Symposium on Image and Signal Processing and Analysis, 316-321, 2005.

[FY01] Feng G.C. Yuen P.C., Multi-cues eye detection on gray intensity image, Pattern Recognition, 34, 1033-1046, 2001.

[FBV02] Féraud R., Bernier O., Viallet J., Collobert M.: A fast and accurate face detector based on neural networks, IEEE Transactions on Pattern Analysis and Machine Intelligence, 23(1), 42-53, 2002.

[FS96] Freund Y. & Shapire R.E., Experiments with a new boosting algorithm, International Conference on Machine Learning, pp 148-156, 1996.

[FS97] Freund Y & Schapire R.E. A decision-theoretic generalization of on-line learning and an application to boosting, Journal of Computer System Sciences, 55:119-139, 1997.

[FR2005] Fumera G.; Roli F., A theoretical and experimental analysis of linear combiners for multiple classifier systems, IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(6), pp 942-956, 2005.

[GMC97] Gader P.D., Mohamed M. & Chiang J.-H., Handwritten Word Recognition with Character and Inter-Character Neural Networks, IEEE Transaction on systems, Man and Cybernetics, 27(1):158–164, 1997.

[G96] Garcia-Salicetti S., Une approche neuronale prédictive pour la reconnaissance en-ligne de l'écriture cursive, Thèse de Doctorat de l'Université Paris VI, 1996.

[GD04] Garcia C. & Delakis M. (2004) Convolutional face finder: A neural architecture for fast and robust face detection, IEEE Transactions on Pattern Analysis and Machine Intelligence, 26(11): 1408-1423.

[GLP07] Geronimo D., Lopez A., Ponsa D. & Sappa A.D., Haar wavelets and edge orientation histograms for on-board pedestrian detection, IbPRIA, pp 418–425, 2007.

[GRB00] Giacinto G., Roli F. & Bruzzone L., Combination of neural and statistical algorithms for supervised classification of remote-sensing images, Pattern Recognition Letters, 21, pp 385-397, 2000.

[GRB07] Grabner H., Roth P.M. & Bischof H., Eigenboosting: combining discriminative and generative information, IEEE Conference on Computer Vision and Pattern Recognition, à paraître, 2007.

[GAL91] Guyon I., Albrecht P., LeCun Y., Denker J. & Hubbard W. Design of a Neural Network Character Recognizer for a Touch Terminal, Pattern Recognition, 24(2): 105-119, 1991.

[GGN06] Feature Extraction, Foundations and Applications, Guyon I, Gunn S., Nikravesh M. & Zadeh L., Editors, Series Studies in Fuzziness and Soft Computing, Physica-Verlag, Springer, 2006.

[GHA92] Guyon I., Henderson D., Albrecht P., Le Cun Y.& Denker J., Writer independent and writer adaptative neural network for on-line character recognition, From Pixel to Features III, 1992.

[GSP94] Guyon I., Schomaker L., Plamondon R., Liberman M. & Janet S. UNIPEN project of on-line data exchange and recognizer benchmarks, International Conference on Pattern Recognition, pp. 29-33, 1994.

[HSD73] Haralick R.M., Shanmugam K. & Dinstein I., Textual Features for Image Classification, IEEE Transactions on Systems, Man, and Cybernetics, 3(6), pp.610-621, 1973.

[HS01] Henning A. & Sherkat N., Cursive script recognition using wildcards and multiple experts, Pattern Analysis & Applications, 4 :51–60, 2001.

[HF91] Higgins C.A. & Ford D.M., Stylus driven interfaces - the electronic paper concept, International Conference on Document Analysis and Recognition, (B), pp 853-862, 1991.

[HAW04] Huang C., Ai H., Wu B. & Lao S., Boosting nested cascade detector for multi-view face detection, International Conference on Pattern Recognition, (2), pp 415-418, 2004.

[HAL07] Huang C., Ai H., Li Y. & Lao S., high-performance rotation invariant multiview face recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence, 29(4): 671-686, 2007.

[HSK05] Huang L. L., Shimizu A. & Kobakate H., Robust face detection using Gabor filter features, Pattern Recognition Letters, 26(11):1641-1649, 2005.

[HS94] Huang Y.S. & Suen C.Y., a method of combining multiple classifiers - a neural network approach, International Conference on Pattern Recognition, (B), pp 473-475 1994.

[HB62] Hubel D. & Wiesel T. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex, Journal of Psychology, 160: 106-154.

[JDM00] Jain A.K., Duin R. & Mao J. Statistical pattern recognition: a review, IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(1): 4-37, 2000.

[KKK05] Kim J.B., Kee S.C. & Kim J.Y., Fast detection of multiview face and eye based on cascaded classifier, Machine vision and Applications, 116-119, 2005.

[K98] Kittler J., Combining Classifiers: A theorical Framework, Pattern Analysis & Applications, 1, pp 18-27, 1998.

[K92] Knerr S., Personnaz L. & Dreyfus G., Une nouvelle approche de la reconnaissance de chiffres manuscrits par réseaux de neurones, Congrès National sur l'Ecrit et le Document, pp 325-332, 1992.

[K89] Kohonen T., Self-organisation and associative memory, Springer-Verlag, 3rd Ed, 1989.

[K07] Krichen E., Reconnaissance de personnes par l'iris en mode dégradé, Thèse de doctorat de l'Institut National des Télécommunications, 2007.

[KW03] Kuncheva L.I. & Whitaker C.J., Measures of diversity in classifier ensembles, Machine Learning, 51, 181-207, 2003.

[LBB98] LeCun Y., Bottou L., Bengio Y. & Haffner P. (1998) Gradient Based Learning Applied to Document Recognition, Proceedings of IEEE, 86(11): 2278-2324.

[LDS90] LeCun Y., Denker J., Solla S., Howard R. E. & Jackel L. D., Optimal brain damage, Advances in Neural Information Processing Systems II, 1990.

[LC07] Leyrit L. & Chateau T., Association de classifieurs pour la sélection de variables, GdR ISIS, Journée détection et reconnaissance d'objet dans les images, 2007.

[L02] Li H., Traitement de la variabilité et développement de systèmes robustes pour la reconnaissance de l'écriture manuscrite en-ligne, Thèse de doctorat de l'Université Pierre et Marie Curie-Paris 6, 2002.

[LZ04] Li S.Z. & Zhang Z., Floatboost learning and statistical face detection, IEEE Transactions on Pattern Analysis and Machine Intelligence, 26(9):1112-1123, 2004.

[LDL05] Liang J., Doermann D. & Li H., Camera-based analysis of text and documents: a survey, International Journal on Document Analysis and Recognition, 7(2-3), pp 84-104, 2005.

[LKP03] Lienhart R., Kuranov A. & Pisarevsky V., Empirical analysis of detection cascades of boosted classifiers for rapid object detection, DAGM03, pp 297–304, 2003.

[L99] Lowe D.G., Object recognition from local scale invariant features, International Conference on Computer Vision, pp 1150–1157, 1999.

[MR81] Mac Lelland J.L. & Rumelhart D.E., An interactive activation model of context effects in letter perception, Psychological Review, Vol. 88, pp 375–407, 1981.

[MFW94] Manke S., Finke M. & Waibel A., Combining Bitmaps with Dynamic Writing Information for On-Line Handwriting Recognition, International Conference on Pattern Recognition, pp 596-598, 1994.

[M93] Milgram M., Reconnaissance des formes. Méthodes numériques et connexionnistes, Armand Colin, 1993.

[MSC05] Milgram J., Sabourin R. & Cheriet M. Combining Model-based and Discriminative Approaches in a Modular Two-stage Classification System: Application to Isolated Handwritten Digit Recognition, Electronic Letters on Computer Vision and Image Analysis, 5(2):1-15, 2005.

[MP97] Moghaddam, B. Pentland, A.: Probabilistic visual learning for object representation, IEEE Transactions on Pattern Analysis and Machine Intelligence, 19(7): 696-710, 1997.

[MA06] Mouchère H. & Anquetil E., Synthèse de caractères manuscrits en-ligne pour la reconnaissance de l'écriture, Colloque International Francophone sur l'Ecrit et le Document, pp 187-192, 2006.

[OS02] Oh I.S. & Suen C.Y.: A class-modular feed-forward neural network for handwriting recognition, Pattern Recognition, 35: 229-244, 2002.

[OSB03] Oliveira, L.S., Sabourin, R., Bortolozzi, F. et Suen, C.Y., A Methodology for Feature Selection Using Multi-Objective Genetic Algorithms for Handwritten Digit String Recognition, the International Journal of Pattern Recognition and Artificial Intelligence, 17(6), pp 903-929, 2003.

[O03] Oudot L., Fusion d'informations et adaptation pour la reconnaissance de textes manuscrits dynamiques, Thèse de doctorat, Université Pierre et Marie Curie-Paris 6, 2003.

[PNM82] Paap K., Newsome S.L., Mac Donald J.E., Schvaneveldt R.W., An activation-verification model for letter and word recognition: The word superiority effect, Psychological Review, Vol. 89, pp 573–594, 1982.

[PI07] Pang Z. & Isbell C., Managing domain knowledge and multiple models with boosting, International Joint Conference on Artificial Intelligence, 2007.

[PP00] Papageorgiou C. & Poggio T., A trainable system for object detection, International Journal of Computer Vision, 38(1): 15–33, 2000.

[P00] Pasquer L., Conception d'un modèle d'interprétation multicontextuelle, application à la reconnaissance en-ligne d'écriture manuscrite », Thèse de Doctorat, Université Rennes I, 2000.

[PCR05] Peng P., Chen, L., Ruan, S., Kukharev, G.: A Robust and Efficient Algorithm for Eye Detection on Gray Intensity Face. International Conference on Advances in Pattern Recognition, Lecture Notes in Computer Sciences, Vol. 3687, 302-308, 2005.

[PC93] Perrone M.P. & Cooper L. N., When Network Disagrees: Ensemble Methods for Hybrid Neural Network, Neural Networks for Speech and Image Processing, Chapman - Hall, pp 126-142, 1993.

[PBC05] Phung S.L., Bouzerdoum A. & Chai D., Skin segmentation using color pixel classification: Analysis and comparison, IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(1), pp 148-154, 2005.

[P91] Plamondon R., Step toward the production of an electronic pen-pad, International Conference on Document Analysis and Recognition, (A), pp 361-371, 1991.

[PS00] Plamondon R. & Srihari S.N. On-line and off-line handwriting recognition: a comprehensive survey, IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(1):63-84, 2000.

[PM97] Platt J.C. & Matt N.P., A constructive RBF network for writer adaptation, Advances in neural information processing systems, Vol. 1, pp. 765–771, 1997.

[PVL04] Poisson E., Viard-Gaudin C. & Lallican P.M., Système TDNN/HMM de reconnaissance de mots cursifs en ligne à apprentissage simplifié, Colloque International Francophone sur l'Ecrit et le Document, 2004.

[PSW97] Powalka R.K., Sherkat N. & Whitrow R.J., Word shape analysis for a hybrid recognition system, Pattern Recognition, 30(3): 421–445, 1997.

[PKP94] Price D., Knerr S., Personnaz L. & Dreyfus G. Pairwise neural network classifiers with probabilistic outputs, Neural Information Processing Systems, 7, 1994.

[PE04] Prudent Y. Ennaji A., La topologie des données pour une distribution fiable des tâches de classification, 11èmes Rencontres de la Société Francophone de Classification, pp 282-285, 2004.

[PNC94] Pudil P., Novovicova J. & Kittler J., Floating search methods in feature selection, Pattern Recognition Letters, 15(11), pp 1119-1125, 994.

[QDM05] Quost B., Denoeux T. & Masson M., Pairwise classifiers in the framework if belief functions, International Conference on Information Fusion, 2005.

[RLE98] Ribert A., Lecourtier Y., Ennaji A. & Stocker E., Vers un classifieur neuronal incrémental : une construction évolutive de taxinomies numériques, Colloque International Francophone sur l'Écrit et le Document, pp 141-150, 1998.

[RSN03] Raina R., Shen Y., Ng A.Y. and McCallum A. Classification with hybrid generative/discriminative models, Neural Information Processing Systems 16, 2003.

[RF03] Rahman A.F. & Fairhust M.C. Multiple classifier decision combination strategies for character recognition: a review, International Journal on Document Analysis and Recognition, 5: 166-194, 2003.

[R03] Ratzlaff E.H., Reports and survey for the comparison of diverse isolated character recognition results on the UNIPEN database, International Conference on Document Analysis and Recognition, (1), pp 623-628, 2003.

[RCB06] Rodriguez Y., Cardinaux F., Bengio S. & Mariethoz J., Measuring the performance of face localization systems, Image and Vision Computing, 24(8): 882:893, 2006.

[RGV01] Roli F., Giacinto G. & Vernazza G., Methods for Designing Multiple Classifier Systems, Multiple Classifier Systems, LNCS 2096, pp 78–87, 2001.

[RBK98] Rowley H.A., Baluja S. & Kanade T., Neural network based face detection, IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(1): 23-38, 1998.

[SL91] Safavian S.R. & Landgrebe D., A survey of decision tree classifier methodology, IEEE Transactions on Systems, Man and Cybernetics, 21(3), pp 660-674, 1991.

[SFB97] Schapire R.E., Freund Y., Bartlett P. & Lee W.S., Boosting the margin: a new explanation for the effectiveness of voting methods, International Conference on Machine Learning, pp 322-330, 1997.

[SK00] Schneiderman & Kanade T., A statistical model for 3D object detection applied to faces and cars, Conference on Computer Vision and Pattern Recognition, (1), pp 746-751, 2000.

[SM96] Schwenk H. & Milgram M. Constraint tangent distance for on-line character recognition, International Conference on Pattern Recognition, (D), pp 520-524, 1996.

[SR98] Senior A. W. & Robinson A. J., An off-line cursive handwriting recognition system, IEEE Transaction on Pattern Analysis and Machine Intelligence, 20(3) :309–321, 1998.

[SRP07] Serre T., Riesenhuber M. & Poggio T., Robust object recognition with cortex-like mechanism, IEEE Transactions on Pattern Analysis and Machine Intelligence, 29(3), 2007.

[SBM05] Sun Z., Bebis G. & Miller R., On-road vehicle detection using evolutionary gabor filter optimization, IEEE Transactions on Intelligent Transportation Systems, 2005.

[SBM06] Sun Z., Bebis G. & Miller R., On-road vehicle detection: A review. IEEE Transactions on Pattern Analysis and Machine Intelligence, 28(5) :694–711, 2006.

[SP98] Sung K.K & Poggio T., Example-based learning for view-based human face detection, IEEE Trans. PAMI, 20(1), pp 39-51, 1998.

[T91] Taft M., Reading and mental lexicon, Erlbaum Edition, 1991.

[TSW90] Tappert C., Suen C. Y. & Wakahara T., The state of the art in on-line handwriting recognition, IEEE Transaction on Pattern Analysis and Machine Intelligence, 12(8):787–808, 1990.

[V95] Vapnik V.N., The nature of statistical learning theory, Springer-Verlag, 1995.

[VMP03] Velázquez R., Maingreaud F. & Pissaloux E., Intelligent Glasses: A New Man-Machine Interface Concept Integrating Computer Vision and Human Tactile Perception, EuroHaptics 2003, Dublin, Ireland, July 2003.

[VJ01] Viola P. & Jones M., Rapid object detection using a boosted cascade of simple features, International Conference on Computer Vision and Pattern Recognition, (1), pp 511-518, 2001.

[VLK02] Vuori V., Laaksonen J. & Kangas,J. Influence of erroneous learning samples on adaptation in on-line handwriting recognition, Pattern Recognition, 35(4): 915-925, 2002.

[VSV03] Vuurpijl L., Schomaker L. & Van Erp M. Architecture for detecting and solving conflicts: two stage classification and support vector classifiers, International Journal on Document Analysis and Recognition, 5: 213-223, 2003.

[WHH89] Waibel A., Hanazawa T., Hinton G., Shikano K. & Lang K., Phoneme recognition using time-delay neural networks, IEEE Transactions on Acoustic Speech &Signal Processing, 37(3), pp 328-339, 1989.

[XKS92] Xu L., Kryzak A. & Suen C.Y., Method of combining multiple classifiers and their application to handwriting recognition, IEEE Transactions on Systems, Man & Cybernetics, 22(3), pp 418-435, 1992.

[YH94] Yang G. & Huang T. S., Human Face Detection in Complex Background, Pattern Recognition, 27(1), pp. 53-63, 1994.

[YKA01] Yang M.H., Kriegman D. & Ahuja N., Face detection using multimodal density models, Computer Vision and Image Understanding, 84(2), pp. 264-284, 2001.

[YKA02] Yang M.H., Kriegman D. & Ahuja N., Detecting Faces in Images: A Survey, IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(1), pp. 34-58, 2002.

[YHC92] Yuille A., Hallinan P. & Cohen D., Feature extraction from faces using deformable templates, International Journal of Computer Vision, 8(2), 99-111, 1992.

[ZCR03] Zhao W., Chellappa R., Rosenfeld A. & Phillips P.J., Face Recognition: A Literature Survey, ACM Computing Surveys, pp. 399-458, 2003.

[ZRZ05] Zhu J., Rosset S., Zhou H. & Hastie T., Multi-class AdaBoost, Technical Report, 2005.

[ZYC06] Zhu Q., Yeh M.C., Cheng K.T. & Avidan S., Fast human detection using a cascade of histograms of oriented gradients, International Conference on Computer Vision and Pattern Recognition, pp 1491–1498, 2006.

7. RECUEIL DE PUBLICATIONS

[Annexe 1] **Prevost L.** & Milgram M., Modelizing character allographs in omni-scriptor frame: a new non-supervised algorithm, **Pattern Recognition Letters**, 21(4), pp 295-302, 2000.

[Annexe 2] **Prevost L.**, Oudot L., Moises A., Michel-Sendis C. & Milgram M., Hybrid generative/discriminative classifier for unconstrained character recognition, **Pattern Recognition Letters**, Special issue on Artificial Neural Networks in Pattern Recognition, 26(12), pp 1840-1848, 2005.

[Annexe 3] **Prevost L.** & Oudot L., Self-supervised adaptation for on-line script text recognition, **Electronic Letters on Computer Vision and Image Analysis**, Special issue on Document Analysis, 5(2), pp 87-97, 2005.

[Annexe 4] Belaroussi R., **Prevost L.** & Milgram M., Algorithm fusion for face localization, **Journal of Advances in Information Fusion**, 1(1), pp 27-38, 2006.

[Annexe 5] Muhammad Hanif S., **Prevost L.**, Belaroussi R. & Milgram M., Real-time facial feature localization by combining space displacement neural networks, **Pattern Recognition Letters**, Special issue on Pattern Recognition in Multidisciplinary Perception and Intelligence, accepté.

[Annexe 6] Negri P., Clady X., Muhammad Hanif S. & **Prevost L.**, A cascade of boosted generative and discriminative classifiers for vehicle detection, **EURASIP Journal on Advances in Signal Processing**, soumis.