Contribution à l'évaluation d'algorithmes de traitement d'images

Christophe Rosenberger

▶ **To cite this version:**

Christophe Rosenberger. Contribution à l'évaluation d'algorithmes de traitement d'images. Interface homme-machine [cs.HC]. Université d'Orléans, 2006. tel-00194404

HAL Id: tel-00194404
https://theses.hal.science/tel-00194404

Submitted on 6 Dec 2007

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

UNIVERSITE D'ORLEANS

HABILITATION A DIRIGER DES RECHERCHES

DE

L'UNIVERSITÉ D'ORLÉANS

Discipline : **Sciences et Technologies Industrielles**

PAR

Christophe Rosenberger

Contribution à l'évaluation

d'algorithmes de traitement d'images.

Soutenue le : **7 Décembre 2006**

MEMBRES DU JURY

- Philippe Bolon	Professeur	Université de Savoie	**Rapporteurs**
- Sylvie Philipp-Foliguet	Professeur	ENSEA - Université de Cergy-Pontoise	
- Marinette Revenu	Professeur	ENSICAEN	
- Atilla Baskurt	Professeur	INSA de Lyon	**Examinateurs**
- Pierre Marché	Professeur	ENSI de Bourges	
- Youssoufi Touré	Professeur	Université d'Orléans	
- Frédéric Truchetet	Professeur	Université de Bourgogne	

A bruno,

REMERCIEMENTS

En tout premier lieu, je tiens à remercier Sylvie Philipp-Foliguet, Marinette Revenu et Philippe Bolon de m'avoir fait le grand honneur d'accepter de rapporter sur cette Habilitation à Diriger des Recherches. Je tiens à remercier Atilla Baskurt et Frédéric Truchetet de leur participation à ce jury.

Ce manuscrit relate des travaux réalisés au sein du Laboratoire de Vision et Robotique (LVR). Aussi, je tiens à remercier Youssoufi Touré le directeur de ce laboratoire et Pierre Marché le responsable de l'équipe SIV. Je tiens à remercier chaleureusement tous les membres du LVR pour leur aide et leur convivialité. Je voudrais marquer ma sympathie à l'ensemble du personnel, des étudiants et des stagiaires de l'ENSI de Bourges qui ont grandement contribué à mon épanouissement.

Je remercie Gérard Poisson et Kacem Chehdi pour leurs conseils avisés concernant la rédaction de ce document. Je tiens à rendre hommage à Hélène Laurent et à Bruno Émile, mes amis, pour leur confiance et pour leur motivation. Un travail dans un manuscrit HDR bénéficie du travail réalisé par des doctorants : Cécile, Sebastien, Anant et Pierre-alain, merci à vous pour le travail que vous avez accompli.

Je tiens à remercier la région Bretagne, la région Centre, l'Union Européenne pour leurs soutiens financiers qui m'ont permis de mener à bien ces travaux de recherche.

"La théorie, c'est quand on sait tout et que rien ne fonctionne.
La pratique, c'est quand tout fonctionne et que personne ne sait pourquoi."
A. Einstein

Résumé

J'ai obtenu mon Doctorat en 1999 sur le thème de la segmentation d'images. Depuis septembre 2000, je suis maître de conférences à l'ENSI de Bourges. J'enseigne des cours de traitement d'images et d'intelligence artificielle au niveau Master. Je suis depuis le 1er septembre 2006, le responsable des relations européennes et internationales au sein de cet établissement. Je suis rattaché au Laboratoire Vision Robotique (UPRES EA 2078) et à l'équipe Signal, Image et Vision. Je suis le responsable du groupe Interprétation au sein de cette équipe regroupant 3 maîtres de conférences, un ATER et 2 doctorants. Mes activités de recherche principales sont la modélisation 3D d'objets, la détection/reconnaissance d'objets et l'évaluation d'algorithmes de traitement d'images. Ce manuscrit concerne ce dernier thème correspondant à mon activité majeure de recherche. J'ai participé au co-encadrement de 4 doctorants dont deux spécifiquement sur ce thème (les deux ayant soutenu) et deux étudiants de Master recherche. J'ai valorisé mes activités de recherche au travers de plusieurs publications : 1 chapitre de livre international, 6 revues internationales, 2 revues nationales, 53 conférences internationales et 10 conférences nationales. Je suis titulaire du Contrat d'encadrement doctoral et de recherche (PEDR) depuis le 1er octobre 2005. J'ai été sollicité pour effectuer de nombreuses expertises (projets ANVAR, référé d'articles et projets ANR). J'ai participé à de nombreux contrats industriels afin de contribuer au budget de fonctionnement du groupe Interprétation.

Ce manuscrit comporte 2 parties. La première correspond à une synthèse de mes activités pédagogiques, administratives et recherche. La seconde présente un résumé des mes activités de recherche dans le domaine de l'évaluation d'algorithmes de traitement d'images. Elle est constituée de trois chapitres. Le premier a pour objectif de positionner la problématique de l'évaluation d'algorithmes de traitement d'images. Les différentes approches utilisées dans la littérature y sont présentées. J'ai développé une méthodologie visant à analyser des approches d'évaluation et à en faciliter leurs conceptions et utilisations. Cette démarche est déclinée au travers de deux algorithmes de traitement d'images. Le troisième chapitre de ce manuscrit présente les travaux sur l'évaluation de résultats de compression d'images ultrasonores dans lesquels j'ai été impliqué. Le thème de l'évaluation de la segmentation d'images que j'ai initié au sein du laboratoire est abordée dans le quatrième chapitre. Je conclus ce document par mon projet de recherche.

Table des matières

Introduction **1**

Partie 1 : Curriculum vitae **5**

1 Présentation générale **7**
1.1 Synthèse . . . 7
1.2 Activités d'enseignement . . . 9
1.3 Activités d'administration . . . 11
1.4 Activités de recherche . . . 11
1.5 Encadrement . . . 15
1.6 Rayonnement scientifique . . . 16
1.7 Publications . . . 19

Partie 2 : L'évaluation d'algorithmes de traitement d'images **27**

2 Problématique de l'évaluation d'algorithmes de traitement d'images **29**
2.1 Introduction . . . 29
2.2 Les méthodologies d'évaluation . . . 31
2.2.1 Évaluation des performances . . . 32
2.2.2 Évaluation par adéquation . . . 32
2.2.3 Évaluation par diagnostic . . . 33
2.3 Les critères d'évaluation . . . 35
2.3.1 Les critères supervisés . . . 36
2.3.1.1 Les données synthétiques . . . 36
2.3.1.2 Le jugement subjectif . . . 37
2.3.2 Les critères non supervisés . . . 39
2.3.3 Les critères hybrides . . . 40
2.4 Conclusion . . . 41

3 Evaluation de la compression d'images ultrasonores **43**
3.1 Contexte . . . 43
3.2 Etat de l'art . . . 46

3.2.1 Cas réversible 46
3.2.2 Cas irréversible 47
3.2.2.1 Critères subjectifs 47
3.2.2.2 Critères objectifs 49
3.2.2.3 Critères hybrides 50
3.2.3 Discussion 52
3.3 Notre contribution 53
3.3.1 Comparaison de méthodologies d'évaluation 53
3.3.1.1 Evaluation subjective à un stimulus 53
3.3.1.2 Evaluation subjective comparative 54
3.3.1.3 Evaluation objective 57
3.3.2 Fusion de critères d'évaluation 61
3.3.2.1 Fusion par combinaison linéaire 62
3.3.2.2 Fusion par apprentissage 66
3.4 Conclusion et perspectives 69

4 Evaluation de la segmentation d'images 73
4.1 Contexte 73
4.2 Définitions de la segmentation 74
4.3 État de l'art 77
4.3.1 Evaluation supervisée 77
4.3.1.1 Évaluation de la détection de frontières 78
4.3.1.2 Évaluation de la segmentation en régions 80
4.3.1.3 Discussion 80
4.3.2 Evaluation non supervisée 80
4.3.2.1 Évaluation de la détection de frontières 81
4.3.2.2 Évaluation de la segmentation en régions 81
4.3.2.3 Discussion 84
4.4 Notre contribution 85
4.4.1 Comparaison de critères d'évaluation 85
4.4.1.1 Evaluation supervisée 85
4.4.1.2 Evaluation non supervisée 92
4.4.1.3 Discussion 95
4.4.2 Fusion de critères d'évaluation 96
4.4.2.1 Critère adaptatif 96
4.4.2.2 Fusion par combinaison linéaire 97
4.4.2.3 Fusion par apprentissage 99
4.4.2.4 Discussion 101
4.4.3 Applications des critères d'évaluation 101
4.4.3.1 Comparaison de méthodes de segmentation 101
4.4.3.2 Evaluation de la qualité de résultats d'images multi-composantes 102
4.4.3.3 Aide au paramétrage de méthodes de segmentation . . . 104

4.4.3.4 Segmentation par optimisation de critères d'évaluation . 105
4.4.3.5 Fusion de résultats de segmentation 106
4.5 Conclusion et perspectives 107

Conclusion et projet de recherche 110

Annexes 119

Article EURASIP Journal on Applied Signal Processing 119

Article IEEE Transactions on Medical Imaging 133

Bibliographie 142

Introduction

Après des études en mathématiques et en informatique à l'université de Rennes I, j'ai décidé de travailler dans le domaine du traitement d'images. J'ai préparé ma thèse entre 1996 et 1999, sur le thème de la segmentation d'images au Laboratoire d'Analyse de Systèmes de Traitement de l'Information (LASTI) de Lannion, qui s'appelle désormais l'équipe de Traitement des Signaux et Images Multi-composantes & Multi-modales (TSI2M). J'ai été scientifique du contingent (dans le cadre de mon service militaire) entre octobre 1999 et juillet 2000 à l'École St Cyr. J'ai travaillé à cette occasion dans un centre de recherche de l'armée (CREC Coetquidan) sur la localisation et reconnaissance de cibles dans une image aérienne.

J'ai été nommé maître de conférences à l'ENSI de Bourges en septembre 2000. L'ENSI de Bourges a été créée en 1997. A mon arrivée, j'ai mis en place tous les cours que j'enseigne actuellement. Les modules dont je suis le responsable sont le traitement d'images, l'intelligence artificielle, le multimédia et sont en relation directe avec mes activités de recherche. En tant qu'enseignant dans une école d'ingénieurs, j'implique des étudiants sur mes projets de recherche ou de transfert de technologie.

Je fais partie du Laboratoire de Vision et Robotique UPRES EA 2078 (LVR) depuis mon arrivée a Bourges. J'appartiens à l'équipe Signal, Image, Vision (SIV) qui est l'une des 3 équipes du LVR (les deux autres concernant la robotique et l'automatique). A mon arrivée, j'ai tout d'abord valorisé mes travaux de recherche de thèse. Durant ma thèse, j'ai pu constaté que l'évaluation de la qualité d'un résultat de segmentation constituait à la fois un problème difficile et un enjeu majeur. Je me suis orienté assez vite dans cette voie de recherche. Deux collègues de mon équipe m'ont rejoint sur ces activités. Leur confiance a renforcé mon envie de poursuivre ces travaux. En 2003, le groupe Interprétation a été officiellement créé au sein de l'équipe SIV, j'en assume la responsabilité depuis cette date.

J'ai développé plusieurs partenariats industriels. Le traitement d'images et la reconnaissance de formes ont beaucoup d'applications. J'ai travaillé par exemple sur le contrôle qualité de cerises par vision ou sur la conception d'un ball-trap électronique (en

repérant le vol du plateau par stéréovision).

J'attache une grande importance à la coopération avec d'autres chercheurs. J'ai ainsi eu l'opportunité de travailler avec différents collègues du laboratoire notamment dans le cadre du projet européen OTELO (mObile Tele-Echography using an ultra-Light rObot). Ma contribution à ce projet a essentiellement concerné l'évaluation de la qualité de résultats de compression d'images ultrasonores (ces travaux sont présentés dans le chapitre 3 de ce manuscrit). Ce fut une expérience enrichissante et je les en remercie pour cela. J'ai développé avec enthousiasme des coopérations nationales et internationales. Ceci permet de bénéficier ou d'apporter une compétence afin d'obtenir des résultats de meilleure qualité ou plus rapidement. La recherche est une aventure humaine où chacun amène son expérience, sa compétence et sa motivation.

Je ne détaillerai pas dans ce document les travaux de co-encadrement de deux thèses de doctorat. La thèse de Pierre-Alain Fayolle que je co-encadre concerne la modélisation d'objets 3D. Elle est réalisée dans le cadre d'une coopération avec le Laboratoire d'Informatique Fondamentale d'Orléans (LIFO). La thèse d'Anant Choksuriwong dont je suis aussi le co-encadrant est réalisée au sein du Groupe Interprétation. Cette thèse concerne la détection et la reconnaissance d'objets et pose les bases des futurs algorithmes d'interprétation de scènes que nous souhaitons développer au sein de ce groupe de recherche. Un des objectifs de cette thèse est la coopération avec l'équipe Robotique du laboratoire sur des applications liées à la robotique autonome.

Tous les chercheurs en traitement des images sont conscients de l'importance de l'évaluation des résultats. Il est en effet essentiel d'accéder à l'efficacité d'une méthode proposée afin notamment de quantifier son apport au vu de l'état de l'art. Le traitement d'images est un domaine attractif car il est souvent aisé même pour un néophyte d'apprécier la qualité du résultat d'un traitement. Dans beaucoup d'articles scientifiques dans le passé, les auteurs se sont contentés de faire appel à l'appréciation visuelle du lecteur d'un article pour mettre en évidence l'apport d'une nouvelle méthode. Cette époque est révolue (ou devrait l'être) et il est maintenant nécessaire de prouver la performance et l'apport d'un nouvel algorithme en utilisant des critères d'évaluation ou en faisant appel à plusieurs experts.

Ce manuscrit a pour titre "Contribution à l'évaluation d'algorithmes de traitement d'images". Ce thème correspond à mes activités principales de recherche et d'encadrement. J'ai deux objectifs au sein de ce thème. Le premier a concerné l'étude des méthodologies existantes pour évaluer la qualité d'un algorithme de traitement d'images. Nous verrons dans ce manuscrit plus spécifiquement l'évaluation de la segmentation et de la compression d'images. L'idée est d'apporter des éléments de réponse à la question que beaucoup se posent : Quel critère ou quelle méthode dois-je utiliser pour quantifier la qualité du résultat d'un algorithme de traitement d'images ? Ce premier objectif a pour but la caractérisation de la performance des méthodes existantes. Pour ce faire,

il est nécessaire de comparer l'efficacité de ces méthodologies dans un cadre le plus rigoureux possible. Le deuxième axe de recherche est la proposition de nouveaux critères d'évaluation permettant d'établir un jugement sur la qualité d'un résultat d'un traitement d'images le plus pertinent possible.

Ce manuscrit se compose de quatre chapitres. Le premier chapitre est consacré à une présentation générale de mes travaux en enseignement, administration et recherche. Le chapitre 2 développe la problématique de l'évaluation d'algorithmes de traitement d'images. Le chapitre 3 correspond aux activités que j'ai menées sur l'évaluation de la compression d'images ultrasonores au travers du co-encadrement de la thèse de Cécile Delgorge. Le chapitre 4 présente les travaux sur l'évaluation de la segmentation d'images. Ils concernent des travaux personnels à l'issue de ma thèse de doctorat et des activités liées au co-encadrement de la thèse de Sébastien Chabrier. Je conclus ce document par mon projet de recherche.

PARTIE I

Curriculum vitae

Chapitre 1

Présentation générale

Ce premier chapitre présente dans un premier temps une synthèse décrivant mon parcours, mes activités pédagogiques, administratives et en recherche. Les activités pédagogiques et administratives sont ensuite détaillées. Concernant les activités de recherche, je met l'accent sur mes thèmes de recherche, mon activité d'encadrement, les différents contrats de recherche auxquels j'ai participé et enfin mes publications.

1.1 Synthèse

Christophe ROSENBERGER
Maître de Conférences à l'ENSI de Bourges (section CNU 61)
Date de naissance : 11 octobre 1973
Situation familiale : marié, 1 enfant
Laboratoire de rattachement : Laboratoire Vision Robotique
Équipe de recherche : Équipe Signal, Image et Vision - Groupe Interprétation
Adresse électronique : christophe.rosenberger@ensi-bourges.fr
Page Web : http ://www.ensi-bourges.fr/perso/rosenberger/

TAB. 1.1 – Etat civil

depuis 2000	**Maître de conférences** à l'ENSI de Bourges - LVR
1996-1999	**DOCTORAT** de l'Université de Rennes 1 - ENSSAT (LASTI)
1993-1996	**MAGISTÈRE** Modélisation Mathématiques et Méthodes Informatiques - Université de Rennes 1.
1995-1996	**DEA** en Informatique - Université de Rennes 1.
1993-1995	**Licence et Maîtrise** de Mathématiques - Université de Rennes 1.

TAB. 1.2 – Parcours

Enseignement	Traitement des images, Intelligence artificielle, Multimédia
Responsabilités	Webmaster de l'ENSI de Bourges (2002 à 2004) Webmaster de l'équipe Signal Image Vision du LVR (depuis 2000) Responsable du groupe Interprétation du LVR (2003) Membre élu au conseil scientifique de l'ENSI de Bourges (2003) Responsable des Relations Européennes et Internationales à l'ENSI de Bourges (2006)

TAB. 1.3 – Activités pédagogiques et administratives

Thèmes de recherche	Evaluation de la qualité d'un algorithme de traitement d'images Détection et reconnaissance d'objets dans une image Modélisation d'objets 3D
Publications	1 chapitre de livre international 6 revues internationales (2 JEI, 1 IEEE ITAB, 1 IEEE TMI, 1 EURASIP, 1 LNCS) 2 revues nationales (Traitement du signal, Hermès) 53 conférences internationales avec actes et comité de lecture 10 conférences nationales avec actes et comité de lecture 5 communications sans comité de lecture et sans actes 8 rapports de recherche.
Rayonnement	Co-organisation de 3 manifestations (réunion GDR ISIS, session spéciale à la conférence CTS et ISSPA) Collaborations nationales (TSI2M, LIFO, PSI) et internationales (USA, Chili, Portugal, Japon, Chine et Canada) Comité de lecture de 10 conférences internationales et 5 revues Expertise de 4 projets pour l'agence nationale de la recherche (ANR) Expertise d'un projet pour l'ANVAR
Encadrement	**Contrat d'encadrement doctoral et de recherche (PEDR)** (2005) 4 thèses (2 soutenues, 2 en cours) 2 stages de Master recherche

TAB. 1.4 – Activités de recherche

1.2 Activités d'enseignement

Dans la mesure où l'ENSI de Bourges fut créée en 1997, à mon arrivée en 2000, il a fallu que je conçoive la majeure partie de mes cours. Ils sont pour la plupart en relation avec mes activités de recherche (voir tableau TAB 1.5).

Module	**Public**	**Volume eq. TD**
Vision assistée par ordinateur	BAC+5	55h
Traitement des images et environnement	BAC+5	27h
Traitement des images et de la vidéo	BAC+4	16h
Compression et tatouage de données multimédia	BAC+4	16h
Intelligence Artificielle	BAC+4	27h
Encadrement de projets	BAC+4	40h

TAB. 1.5 – Enseignements dispensés à l'ENSI de Bourges

∗ **Vision assistée par ordinateur**
Public : 3ème année d'école d'ingénieurs (filière MRI options TPR et RSI)
Volume horaire : 10h40 CM, 10h40 TD, 40h TP
Objectif : De par leur modularité et leur faible coût relatif, les systèmes de vision deviennent un outil indispensable du contrôle des systèmes industriels. Ce module offre les bases fondamentales de la vision (modèle de caméra, stéréovision, analyse d'images, compression d'mages). Les travaux pratiques traitent notamment des applications de la vision pour le contrôle qualité ou la vidéosurveillance.

∗ **Traitement des images et environnement**
Public : 3ème année d'école d'ingénieurs (filière MRI option RE)
Volume horaire : 10h40 CM, 10h40 TD
Objectif : Une multitude de capteurs (infrarouge, satellitaires...) permettent d'avoir des informations sur un phénomène lié à l'environnement (pollution de l'environnement, catastrophes naturelles...) sous la forme d'une image. Le traitement d'images permet alors d'automatiser l'extraction d'informations dans une image. Des illustrations sur des cas concrets liés à l'environnement sont présentées (détection de déversement d'hydrocarbures dans l'océan, impact d'une tornade dans une région...).

∗ **Traitement des images et de la vidéo**
Public : 2ème année d'école d'ingénieurs (filière STI dominante multimédia)
Volume horaire : 10h40 CM
Objectif : Ce module offre les bases fondamentales du traitement d'images (acquisition, pré-traitements, segmentation, interprétation) et de la vidéo (détection et estimation de mouvement). Les différents traitements sont présentés en cours et mis en application durant des travaux pratiques.

∗ **Tatouage et compression de données multimédia**
Public : 2ème année d'école d'ingénieurs (filière STI dominante multimédia)
Volume horaire : 10h40 CM
Objectif : La transmission de données multimédia pose deux problèmes. Le premier concerne leur taille importante, une phase de compression est donc nécessaire. Le second concerne la propriété des données. Les standards de compression sont alors présentés et différentes techniques de tatouage développées.

∗ **Intelligence artificielle**
Public : 2ème année d'école d'ingénieurs (filière MRI dominante SA)
Volume horaire : 10h40 CM, 10h40 TD
Objectif : L'intelligence artificielle est un module regroupant un ensemble de techniques visant à reproduire par un programme informatique des aptitudes dites "intelligentes". On peut citer la capacité à raisonner (logique), à apprendre de nouvelles connaissances (apprentissage), à décomposer un problème... Ces capacités permettent de remplacer ou de faciliter la prise de décision d'un humain pour l'automatisation d'un système.

Projets

J'ai souvent des étudiants en projet sur des applications de transfert technologique pour un industriel ou sur un projet de recherche. Dans le cas de projets de transfert de technologie, les étudiants sont en contact direct avec les industriels (définition du cahier des charges, réunions et états d'avancement). Un exemple d'un tel projet est le système que nous avons conçu pour le contrôle qualité de cerises par vision. Par ailleurs, je propose souvent des sujets avec un côté ludique comme par exemple le projet du robot qui joue au billard (voir figure FIG 1.1).

FIG. 1.1 – Exemple de projet étudiant encadré : Le robot qui joue au billard

1.3 Activités d'administration

Responsabilités au sein de l'ENSI de Bourges

L'ENSI de Bourges compte actuellement presque 400 étudiants et 23 enseignants titulaires. Dans ce contexte, les enseignants occupent différentes responsabilités pour le bon fonctionnement de l'établissement. Ainsi, j'ai été de 2000 à 2002, le coordinateur de la matière Cindynique (maîtrise des risques) où 15 vacataires professionnels interviennent. De 2002 à 2004, j'ai occupé la fonction de webmaster de l'école. Depuis septembre 2006, j'occupe la responsabilité des relations européennes et internationales à l'ENSI de Bourges. Je suis depuis 2003 membre élu du corps des maîtres de conférences au conseil scientifique de l'école.

Responsabilités au sein LVR

L'équipe SIV est l'une des trois équipes du laboratoire. Elle regroupe 9 enseignants-chercheurs titulaires (1 PR et 8 MCF). La thématique de recherche de cette équipe concerne le développement d'outils d'aide à la décision à partir de systèmes de perception. Deux groupes de recherche dans cette équipe s'intéressent à des thèmes plus spécifiques. Le premier groupe intitulé « Vision » s'intéresse d'une part, à la reconstruction et à l'analyse de formes 3D (diagnostic sur relevés anatomiques, détection d'obstacles dans des scènes routières...) et d'autre part, à la conception et au réglage des systèmes de vision. J'assure la responsabilité du second groupe appelé « Interprétation » depuis sa création en 2003. Le terme interprétation fait référence d'une part, à l'interpretation de résultats de traitement d'images (évaluation de performance) et d'autre part, au développement de systèmes d'interprétation d'images. Ce groupe est composé actuellement de 3 enseignants chercheurs, d'un ATER et de deux doctorants. Nous avons un rythme de réunion mensuelle permettant un suivi régulier de l'avancement des thèses et activités de recherche du groupe. Je m'occupe depuis 2000 de la mise à jour de page Web de l'équipe SIV du LVR. J'administre aussi une salle du laboratoire au sein de l'ENSI de Bourges que j'ai complètement montée. Elle comporte environ 8 ordinateurs et différents matériels de vision.

1.4 Activités de recherche

- **1996-1999** - Laboratoire d'Analyse des Systèmes de Traitement de l'Information (LASTI - UPRES EA 1252) s'appellant désormais l'équipe de Traitement des Signaux et Images Multi-composantes & Multi-modales (TSI2M) : thèse de doctorat sous la direction de K. Chehdi.

  Le cadre général de ma thèse était celui de la segmentation d'images. J'ai développé un système adaptatif de segmentation qui a été appliqué notamment sur des images pour des applications en télédétection. J'ai abordé dans la thèse le

problème lié à l'évaluation de la qualité d'un résultat de segmentation, la fusion et la segmentation d'images multi-composantes.

Après un travail de synthèse permettant de répertorier les différentes méthodes de segmentation existantes en fonction du type d'informations pour lesquelles elles sont performantes, un système original de segmentation a été proposé. L'originalité de ce système résidait dans l'adaptation des traitements au contexte local de l'image avec le minimum de connaissances *a priori*. Il est constitué de trois modules de traitement. *Le premier module* permet d'analyser finement l'image à deux niveaux. Le premier niveau identifie d'une part, le contexte global de l'image à traiter (image majoritairement composée de régions uniformes ou texturées) afin d'adapter la suite des traitements et, distingue d'autre part, les zones texturées et uniformes la composant. Le second niveau du module concerne l'analyse locale de l'image à segmenter afin de caractériser chacune des régions détectées par des attributs classiques de textures pertinents (obtenus par analyse statistique) et des attributs que nous avons définis. Ces paramètres complémentaires ont été déterminés à partir d'un modèle de texture basé sur la décomposition de Wold de la fonction d'autocovariance. Ils permettent d'obtenir des informations sur le type de texture (aléatoire ou déterministe) et sur sa granularité (grossière ou fine). Cette analyse plus fine d'une région texturée permet, d'une part, de faciliter le choix de la méthode de segmentation appropriée et, d'autre part, d'adapter la taille du support d'analyse de la région à segmenter. *Le deuxième module* déclenche la méthode de segmentation adaptée au contexte local de l'image en utilisant une méthode de classification non supervisée que nous avons développée. *Le troisième module* permet de fusionner, soit les résultats de plusieurs méthodes de segmentation d'une même image, soit les résultats de segmentation de chaque bande dans le cas d'une image multi-composantes. La méthode de fusion développée adopte une approche génétique en combinant les résultats de segmentation pondérés par un critère d'évaluation proposé dans le cadre de la thèse. Le système a été validé sur différents types d'images, synthétiques et réelles (télédétection).

Mots-Clefs : Analyse d'images, Segmentation adaptative, Analyse de textures, Fusion de résultats de segmentation, Algorithme génétique, Evaluation de résultats de segmentation, Méthode de classification automatique, Segmentation multi-composantes.

Jury : C. Fernandez-Maloigne, A. Hillion (Rapporteurs), J.M. Chassery (President), K. Chehdi, I.C. Lerman, S. Philipp-Foliguet (Examinateurs).

- **1999-2000** - Centre d'Etude et de Recherche des Écoles de Coetquidan (CREC) : scientifique du contingent (service militaire) sous la responsabilité du lieutenant Cozien.

Le travail à réaliser entrait dans le cadre du Projet JAROD développé au sein du CREC. Ce projet avait pour but le développement d'un système automatique de reconnaissance de cibles (avions posés sur un aéroport) à partir d'images aériennes et satellitaires. L'originalité de ce système résidait dans sa conception combinant une approche multi-agents et neuronale et dans le fait qu'il constituait un réseau distribué asynchrone. Les agents avaient pour but de segmenter des cibles potentielles et de les fournir à un serveur de réseaux de neurones (pouvant être situé sur une autre machine). Le réseau de neurones, s'il y avait reconnaissance, identifiait la cible et en informait l'agent concerné.

Mon travail a concerné la détection d'objets par un système multi-agents.

Mots-Clefs : Détection de contours, Multi-agents, Réseaux de neurones, Interface, Communication TCP/IP, Génie logiciel

- **depuis 2000** - Laboratoire Vision Robotique (LVR) UPRES EA 2078 : maître de conférences dans l'équipe Signal, Image et Vision.

  Depuis 2000, mes activités de recherche s'articulent autour de trois thèmes :

  a) Evaluation de la qualité d'algorithmes de traitement d'images
  L'objectif de ce thème est l'étude et la définition de critères d'évaluation pour quantifier la qualité d'algorithmes de traitement d'images. La segmentation est une étape essentielle en traitement d'images et reste un problème complexe. Il existe une multitude de méthodes de segmentation dont l'efficacité reste difficile à évaluer. Notre objectif est de contribuer à la comparaison des critères existants et au développement de nouveaux critères plus efficaces. Des travaux sur l'évaluation de la compression d'images ultrasonores ont aussi été réalisés dans le cadre du projet européen OTELO.

  Mots-Clefs : Critères statistiques, Evaluation subjective, Fusion, Apprentissage, benchmark.

  b) Interprétation de scènes pour la robotique
  Ce thème concerne le développement d'une plateforme d'interprétation de scènes. Des descripteurs invariants locaux (moments de Zernike) couplés à une méthode d'apprentissage (séparateur à vaste marge) permettent la détection et la reconnaissance d'objets connus dans une image (voir figure FIG. 1.2). Les applications visées sont notamment l'interprétation de scènes pour la robotique (cela permettra de favoriser une coopération avec l'équipe Robotique du LVR).

  Mots-Clefs : Reconnaissance de formes, Descripteurs invariants, Apprentissage.

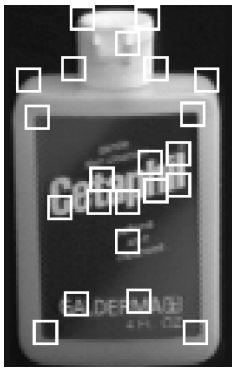


FIG. 1.2 – Description d'un objet par des invariants locaux.

c) Reconstruction 3D d'objets par des fonctions mathématiques
Ce projet concerne la modélisation 3D d'objets numérisés par un scanner (voir figure FIG. 1.3). L'objectif est de modéliser un nuage de points acquis, par un modèle mathématique appelé Frep (Functional representation) pour adapter son rendu. Les travaux ont concerné d'une part la détermination des paramètres du modèle d'un objet. D'autre part, la définition automatique du modèle d'un objet est à l'étude. Ce travail est réalisé en coopération avec le laboratoire d'Informatique Fondamentale d'Orléans (LIFO) et l'université d'Aizu au Japon (Professeur Pasko). Depuis 2005, le groupe ECOS de l'université de Coïmbra participe à ce projet grâce à ses compétences en programmation génétique (pour la partie détermination automatique du modèle).

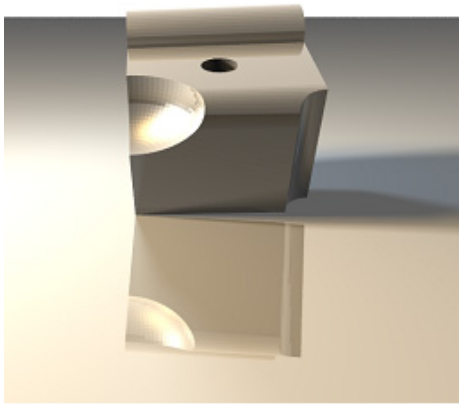


FIG. 1.3 – Exemple d'un objet 3D modélisé.

Mots-Clefs : Reconstruction 3D, Modélisation par des Frep, Algorithmes évolutionnaires.

1.5 Encadrement

Thèses co-encadrées

1. Cécile Delgorge
"Proposition et évaluation de techniques de compression d'images ultrasonores dans le cadre d'une télé-échographie robotisée"
Thèse de Doctorat de l'Université d'Orléans **soutenue le 10 décembre 2005**
co-encadrement (33%) avec Pierre Vieyres et Gérard Poisson.

2. Sébastien Chabrier
"Contribution à l'évaluation des performances de la segmentation d'images"
Thèse de Doctorat de l'Université d'Orléans **soutenue le 14 décembre 2005**
co-encadrement (50%) avec Pierre Marché et Hélène Laurent.

3. Pierre-Alain Fayolle
"Modélisation d'objets 3D par une représentation fonctionelle"
Thèse de Doctorat de l'Université d'Orléans débutée en septembre 2002 (soutenance prévue fin 2006)
co-encadrement (80%) avec Christian Toinard.

4. Anant Choksuriwong
"Interprétation de scènes pour la robotique"
Thèse de Doctorat de l'Université d'Orléans débutée en décembre 2003 (soutenance prévue en 2007)
co-encadrement (50%) avec Pierre Marché et Bruno Emile.

Encadrement de stages de Master Recherche

1. Gilles Bougenière
"Réalisation d'un démonstrateur de ball-trap électronique"
Master Recherche d'Informatique de l'Université d'Orléans
encadrement (100%) de février à août 2005.

2. Baptiste Hemery
"Reconnaissance faciale pour l'authentification biométrique"
Master Recherche d'Informatique de l'Université d'Orléans
co-encadrement (50%) avec Christian Toinard de février à août 2006.

1.6 Rayonnement scientifique

Collaborations nationales

Je participe au Groupement de recherche PRC ISIS sur le thème B (Image et Vision). Je suis impliqué dans le projet ROBIN (détection et reconnaissance d'objets dans des images) du projet national TECHNOVISION.

Je collabore avec différents laboratoires de recherche nationaux :

- L'équipe Traitement des Signaux et Images Multi-composantes & Multi-modales (TSI2M) : j'ai continué de travailler en collaboration avec le laboratoire où j'ai passé ma thèse (6 publications en commun depuis la thèse),
- Le laboratoire Perception, Systèmes et Informations (PSI) FRE CNRS 2645 : je travaille avec Alain Rakotomamonjy, membre de ce laboratoire (ayant effectué sa thèse au sein du LVR) sur l'application des séparateurs à vaste marge (SVM) en traitement d'images (4 publications en commun),
- Le laboratoire d'Informatique Fondamentale d'Orléans (LIFO) : un travail en coopération est mené avec ce laboratoire sur un sujet transversal. Il concerne la reconstruction 3D d'objets (6 publications en commun).

Collaborations internationales

Dans le cadre de mes travaux de recherche, j'ai collaboré avec différents laboratoires internationaux :

- Le groupe de recherche "Evolutionary and Complex Systems" (ECOS) de l'Université de Coïmbra au Portugal : je travaille avec Sara Silva (doctorante) sur la reconstruction 3D d'objets par programmation génétique (2 publications en commun).
- L'Université de Montréal au Canada : je travaille avec Pierre-marc Jodoin (doctorant) sur la fusion d'informations avec un résultat de segmentation (1 publication en commun).
- L'Université de Aizu au Japon : je travaille avec le Pr. Alexander Pasko sur la modélisation d'objets 3D (1 publication en commun).
- L'Université de Tsinghua en Chine : je travaille avec le Pr. Yu-Jin Zhang sur l'évaluation de la segmentation. Il a participé au jury de thèse de Sébastien Chabrier en décembre 2005 à l'ENSI de Bourges (2 publications en commun).
- L'école "Electrical and Computer Engineering Department School of Engineering" de l'Université de Dayton : depuis le séjour de Waleed Smari comme professeur invité à l'ENSI de Bourges en juillet 2004, des liens se sont tissés notamment sur l'utilisation de systèmes multi-agents en traitement d'images (4 publications en commun).
- L'université de Talca au Chili : je travaille avec le Pr. Per Bro sur le contrôle qualité par vision dans le domaine agro-alimentaire. Per Bro a été un professeur invité en

juillet 2005 à l'ENSI de Bourges (5 publications en commun).

J'ai mis en place une convention de partenariat entre l'ENSI de Bourges et ces deux dernières universités pour faciliter l'échange d'étudiants et d'enseignants-chercheurs.

Contrats de recherche

Le LVR dispose d'un budget de fonctionnement limité (dotation ministérielle). L'équipe SIV composée de 9 titulaires dispose ainsi d'un budget de fonctionnement de 6400 € par an. Dans ce contexte, il est nécessaire de trouver des fonds ailleurs afin notamment de financer du matériel et le déplacement à des conférences. J'ai donc travaillé sur plusieurs projets avec des partenaires industriels.

- Le premier projet en 2002 portait sur de la sécurité informatique avec un aspect lié au **tatouage d'images** (confidentiel). Je suis intervenu dans ce projet sur le développement d'une partie spécifique du logiciel. La durée du projet était d'un an et le montant perçu de 1500 €,
- En juin 2002, j'ai obtenu une **Prestation Technologique Réseau** de 8000 € avec une société locale (Calibreuses Fachaux) pour une durée d'un an. J'étais le responsable de ce contrat et j'ai encadré plusieurs étudiants en projet et en stage à cette occasion. Il visait le développement d'un prototype de système de vision permettant de réaliser en temps-réel le contrôle qualité de cerises (calibre, couleur de la cerise et détection de défauts). Ce travail a conduit a plusieurs **publications (1 conférence nationale, 2 conférences internationales et 1 revue internationale)**,
- En mai 2002, j'ai participé à une demande de **Bonus Qualité Recherche (BQR)** à l'Université d'Orléans avec un membre de l'équipe Robotique du LVR. Cela nous a permis de bénéficier pour nos travaux d'une enveloppe de 9000 € pour une durée de 3 ans,
- En septembre 2002, j'ai participé au montage du dossier de subvention du Conseil Général du Cher sur le thème de la reconstruction 3D. Nous avons obtenu une subvention de 50000 € visant à soutenir les travaux de recherche de la thèse de Pierre-Alain Fayolle pendant 3 ans,
- En septembre 2004, nous avons obtenu une subvention de l'ARITT Centre de 10000 € pour l'étude de faisabilité d'un ball-trap électronique sur une période d'un an. L'idée est de concevoir un système reproduisant la sensation d'un ball-trap réel sans l'utilisation d'une cartouche réelle. Un système a été développé utilisant un capteur de localisation de la direction du fusil et une modélisation de la trajectoire du plateau. Les paramètres du modèle de la trajectoire du plateau ont été déterminés par stéréovision. Ce travail a conduit à une publication à 1 conférence internationale,
- En février 2006, j'ai participé au montage d'**un projet du pôle de compétitivité "Sciences et Système de l'Énergie Électrique" (S2E2)**. Le projet CAPTHOM a pour objectif la conception d'un capteur de présence humaine universel. Il permettra d'adapter la consommation d'énergie électrique en fonction de la présence d'un humain (éclairage, chauffage). Des applications en sécurité (détection d'intrusion)

sont aussi envisagées. Le système CAPTHOM qui sera développé utilisera sans doute plusieurs types de capteurs (audio, caméra,...). Une fusion des informations issues de ces capteurs sera réalisée afin d'augmenter la confiance dans la détection. La solution devra être robuste et bon marché afin d'être intégrée lors de la construction future d'habitations particulières. Le montant obtenu (420 000 €) permettra de financer 3 thèses de doctorat et 1 post-doc à la rentrée 2006 pour le laboratoire LVR sur une période de 3 ans. Je vais être impliqué sur la thèse s'intéressant à la modalité image (information issue d'une caméra). Des informations plus haut niveau sur l'individu détecté pourront être exploitées pour des applications d'aide au maintien à domicile de personnes handicapées ou seniors. Il sera possible par exemple de recueillir des informations sur l'activité de la personne ou de détecter un comportement anormal comme par exemple une chute.

Le tableau TAB. 1.6 présente de façon synthétique les financements obtenus.

Type de financement	**Durée**	**Montant**
Contrat transfert technologique (tatouage d'images)	1 an	1 500 €
Contrat transfert technologique (contrôle qualité de cerises)	1 an	8 000 €
Bonus Qualité Recherche	3 ans	9 000 €
Subvention Conseil Général du Cher	3 ans	50 000 €
ARITT Centre	1 an	10 000 €
Projet CAPTHOM (pôle de compétitivité S2E2)	3 ans	420 000 €

TAB. 1.6 – Financements obtenus

Référé d'articles

- participation au comité de lecture des conférences internationales :
  - International Conference on Parallel and Distributed Computing and Systems (PDCS'06),
  - International Conference on Cybernetics and Information Technologies, Systems and Applications (CITSA'06),
  - International Conference of the IEEE Industrial Electronics Society (IECON'06),
  - International Symposium on Industrial Electronics (ISIE'06),
  - Session spéciale "Multi-Agent based Systems and Collaboration (MASC'06)" International Symposium on Collaborative Technologies and Systems (CTS'06),
  - International Conference on High Performance Computing and Simulation (HPCS'05-06),
  - European Conference on Modeling and Simulation (ESM'05),
  - IEEE International Conference on Information Reuse and Integration, (IRIR'04),
  - IEEE International Global Communications Conference, (GlobalCom'04),
  - International Symposium on Signal Processing and its Applications, (ISSPA'03) et (ISSPA'07),

  - International Symposium on Measurement and Control in Robotics, (ISMCR'02).
- participation au comité de lecture des revues :
  - Revue Traitement du Signal,
  - Revue Internationale on Remote Sensing of Environment,
  - Revue Internationale EURASIP on Applied Signal Processing,
  - Revue Internationale on Computer and their Applications,
  - Revue Internationale Electronic Letters on Computer Vision and Image Analysis (ELCVIA).

Organisation de manifestations

- co-organisateur de la journée du GDR ISIS du 15/12/2005 sur le thème de "l'évaluation des algorithmes de traitements d'images",
- co-organisateur de la session spéciale "Multi-Agent based Systems and Collaboration (MASC'06)" lors de la conférence internationale (CTS'06),
- co-organisateur de la session spéciale "Performance evaluation and Benchmarking of Image and Video Processing" lors de la conférence internationale (ISSPA'07).

1.7 Publications

Chapitre de livre international

1. **C. Rosenberger**, S. Chabrier, H. Laurent, B. Emile "Unsupervised and supervised image segmentation evaluation", pages 365-393, chapitre 18 du livre "Advances in Image and Video Segmentation", Pr. Yu-Jin Zhang, Idea Group Publishing, 2006.

Revues internationales à comité de lecture

1. C. Delgorge, **C. Rosenberger**, G. Poisson, P. Vieyres : "Towards a new tool for the evaluation of the quality of ultrasound compressed images", IEEE Transactions on Medical Imaging, Special Issue on Validation in Medical Image Processing, 2006.
2. S. Chabrier, B. Emile, **C. Rosenberger**, H. Laurent : "Unsupervised performance evaluation of image segmentation", Special Issue on Performance Evaluation in Image Processing, EURASIP Journal on Applied Signal Processing, pages 1-12, 2006.
3. C. Delgorge, F. Courreges, L. Al Bassit, C. Novales, **C. Rosenberger**, N. Smith-Guerin, C. Brù, R.Gilabert, M. Vannoni, G. Poisson, P. Vieyres : "A tele-operated mobile ultrasound scanner using a light weight robot", IEEE Transactions on Information Technology in Biomedicine (ITAB), Special Issue mHealth, vol. 9, no.1, mars 2005.

4. S. Silva, P.-A. Fayolle, J. Vincent, G. Pauron, **C. Rosenberger**, C. Toinard : "Genetic Algorithms For Shape Modelling and Fitting", Lecture Notes in Computer Science, ISBN : 3-540-30737-0, vol 3808, pages 144-155, (article sélectionné du workshop ALEA de la Conference EPIA), 2005.
5. **C. Rosenberger**, B. Emile, H. Laurent : "Calibration and Quality Control of Cherries by Artificial Vision", International Journal of Electronic Imaging, Special Issue on quality control by artificial vision, pages 539-546, (13), n°3, Juillet 2004.
6. **C. Rosenberger** , K. Chehdi : "Towards a complete adaptive analysis of an image", International Journal of Electronic Imaging, vol. 12, Issue 2, pp. 292-298, 2003.

Revues nationales à comité de lecture

1. **C. Rosenberger**, K. Chehdi : "Nouveaux attributs de texture invariants par rotation", Revue Traitement du signal, vol. 18, n°5-6, pp 447-454, 2001.
2. K. Chehdi, **C. Rosenberger**, C. Kermad : "Segmentation d'images multi-composantes par fusion génétique", Images et vidéo, Réseaux et Systèmes Répartis, vol. 12, n°3-4, pp. 471-480, Editions Hermès, 2000.

Conférences internationales avec comité de lecture et avec actes

1. S. Chabrier, **C. Rosenberger**, B. Emile : "Evaluation methodologies of image processing : an overview", IEEE International Conference on Signal Processing (ICSP), 2006.
2. S. Chabrier, **C. Rosenberger**, H. Laurent, Y.J. Zhang : "Adaptive evaluation of image segmentation results", IEEE International Conference on Pattern Recognition (ICPR), 2006.
3. B. Hemery, **C. Rosenberger**, C. Toinard, B. Emile : "Comparative study of invariant descriptors for face recognition", IEEE International Conference on Signal Processing (ICSP), 2006.
4. S. Chabrier, H. Laurent, **C. Rosenberger**, Y.J. Zhang : "Supervised evaluation of synthetic and real contour segmentation results", EUSIPCO, 2006.
5. R. Chamelat, E. Rosso, A. Choksuriwong, **C. Rosenberger**, H. Laurent, P. Bro : "Use of Zernike Moments for Grape Detection with Image Processing", Conference of the IEEE Industrial Electronics Society (IECON), 2006.
6. A. Choksuriwong, **C. Rosenberger**, B. Emile : "Parametrization of a local invariant object recognition technique", Conference of the IEEE Industrial Electronics Society (IECON), 2006.
7. P. Bro, **C. Rosenberger**, H. Laurent, C. Gaete-Eastman, M. Fernández, M.A. Moya-Leon : "A support vector machine as an estimator of mountain papaya ripeness using resonant frequency or frequency centroid", in IFIP International Federation for Information Processing, Colume 217, Artificial Intelligence in Theory and Practice, ed. M. Bramer, (Boston :Springer), pages 335-344, 2006.

8. P.-A. Fayolle, S.Silva, G. Latinier, D. Saffrey, **C. Rosenberger**, C. Toinard : "Shape Modeling With Genetic Programming", IEEE Virtual Reality International Conference (VRIC), pp. 235-241, 2006.
9. V. Arguenon, A. Bergues-Lagarde, **C. Rosenberger**, P. Bro, W. Smari : "Multi-Agents Prototyping of Agriculture Robots", Special Session on Multi-Agent based Systems and Collaboration (CTS), 2006.
10. P.-A. Fayolle, A. Pasko, N. Mirenkov, **C. Rosenberger**, C. Toinard : "Constructive Tree Recovery Using Genetic Algorithms", International Conference on Visualization, Imaging and Image processing (VIIP), 2006.
11. G. Bougeniere, P. Moulon, **C. Rosenberger**, W. Smari : "On the determination of 3D Trajectory of Moving Targets by Stereovision", IEEE International Conference on Computers, Communications, and Signal Processing (CCSP), 2005.
12. V. Deligeon, G. Mahe, **C. Rosenberger**, P. Bro, and W. Smari : "Distributed and Scalable Vision System for Quality Control of Cherries", International Conference on Computer Applications in Industry and Engineering (CAINE), 2005.
13. P.B. Bro, **C. Rosenberger**, H. Laurent, C. Gaete, M. Fernandez : "Application of a support vector machine to spectrograms for determination of papaya ripeness", Workshop on Artificial Intelligence, Valdivia, Chile, 2005.
14. A. Choksuriwong, H. Laurent, **C. Rosenberger**, C. Maaoui : "Object recognition using local characterisation and Zernike moments", Lecture Notes in Computer Science, ISBN : 3-540-29032, International Conference on Advanced Concepts for Intelligent Vision Systems (ACIVS), pages 108-115, 2005.
15. S. Chabrier, B. Emile, **C. Rosenberger** : "Texture detection for image analysis", International Conference on Advances in Pattern Recognition (ICAPR), Lecture Notes in Computer science, ISBN 0302-9743, vol 2, pages 455-463, 2005.
16. C. Delgorge, **C. Rosenberger**, G. Poisson, P. Vieyres : "Evaluation of the Quality of Ultrasound Image Compression by Fusion of Criteria with a Genetic Algorithm", International Conference on Advances in Pattern Recognition (ICAPR), Lecture Notes in Computer science, ISBN 0302-9743, vol 2, pages 464-472, 2005.
17. S. Chabrier, **C. Rosenberger**, H. Laurent, A. Rakotomamonjy : "Segmentation Evaluation Using A Support Vector Machine", International Conference on Advances in Pattern Recognition (ICAPR), Lecture Notes in Computer science, ISBN 0302-9743, vol 1, pages 426-435, 2005.
18. C. Delgorge, **C. Rosenberger**, G. Poisson, P. Vieyres : "Evaluation of the Quality of Ultrasound Image Compression for a Robotic Tele-Echographic System", IEEE International Conference on Computers in Cardiology (CINC05), Lyon, 2005.
19. C. Maaoui, H. Laurent, **C. Rosenberger** : "2D shape recognition using Zernike moments", IEEE International Conference on Image Processing (ICIP), 2005.
20. C. Delgorge, **C. Rosenberger**, A. Rakotomamonjy, G. Poisson, P. Vieyres : "Evaluation of the Quality of Ultrasound Image Compression by Fusion of Criteria with a Support Vector Machine", EUSIPCO, 2005.

21. S. Chabrier, **C. Rosenberger**, B. Emile : "Segmentation Evaluation by Fusion with a Genetic Algorithm", EUSIPCO, 2005.
22. C. Maaoui, **C. Rosenberger**, B. Emile : "Robust color object detection and recognition", EUSIPCO, 2005.
23. A. Choksuriwong, **C. Rosenberger**, W. Smari : "Multi-agents System For Image Understanding", IEEE International Conference on Integration of Knowledge Intensive Multi-Agent Systems (KIMAS) Massachusetts, USA, 2005.
24. S. Chabrier, B. Emile, H. Laurent, **C. Rosenberger**, P. Marché : "Unsupervised Evaluation of Image Segmentation : Application to multi-spectral images", IEEE International Conference on Pattern Recognition (ICPR), (3), pages 576-579, Cambridge, 2004.
25. P.-A. Fayolle, **C. Rosenberger**, C. Toinard : "3D Shape Reconstruction of Template Models Using Genetic Algorithms", IEEE International Conference on Pattern Recognition (ICPR), (2), pages 269-272, Cambridge, 2004.
26. P.-A. Fayolle, **C. Rosenberger**, C. Toinard : "Shape Recovery and Functional Modeling Using Genetic Algorithms", IEEE Virtual Reality International Conference (VRIC), pages 227-232, Laval, 2004.
27. S. Chabrier, **C. Rosenberger**, H. Laurent, B. Emile, P. Marché : "Evaluating the segmentation result of a gray-level image", EUSIPCO, pages 953-956, Vienne, 2004.
28. S. Chabrier, H. Laurent, B. Emile, **C. Rosenberger**, P. Marché : "A comparative study of supervised evaluation criteria for image segmentation", EUSIPCO, pages 1143-1146, Vienne, 2004.
29. **C. Rosenberger**, A. Rakotomamonjy, B. Emile : "Generic target recognition", EUSIPCO, pages 1613-1616, Vienne, 2004.
30. **C. Rosenberger**, K. Chehdi : "Unsupervised segmentation of multi-spectral images", International Conference on Advanced Concepts for Intelligent Vision Systems (ACIVS), September 2-5, 2003, Ghent, Belgium.
31. F. Poupeau, **C. Rosenberger**, H. Laurent, B. Emile : "Calibration and Quality Control of Cherries by Artificial Vision", International Conference on Quality Control by Artificial Vision (QCAV), Salt lake city, 2003.
32. **C. Rosenberger**, K. Chehdi : "Supervised Genetic Image Segmentation", IEEE International Conference on System, Man and Cybernetics (SMC), Hammet, 2002.
33. A. Benali, **C. Rosenberger**, P. Marché : "Application of Telerobotic Computer Vision System in Manufacturing Control Process", IEEE International Conference on System, Man and Cybernetics (SMC), Hammamet, 2002.
34. C. Delgorge, **C. Rosenberger**, P. Vieyres, G. Poisson : "JPEG 2000, an adapted compression method for ultrasound images ? A comparative study", International Conference on Systemics, Cybernetics and Informatics (SCI), vol. 9, pp. 536-539, Orlando, 2002.

35. A. Benali, **C. Rosenberger**, P. Marché : "Teleoperation of an Industrial Vision System", International Symposium on Advanced Robot Systems and Virtual Reality (ISMCR), ss :B8, Bourges 2002.
36. **C. Rosenberger**, C. Cariou : "Contribution to Texture Analysis", International Conference on Quality Control by Artificial Vision (QCAV) ,vol. 1, pp. 122-126, Le Creuzot, 2001.
37. C. Delgorge, P. Vieyres, G. Poisson, **C. Rosenberger**, P. Arbeille : "Comparative Survey of Ultrasound Images Compression Methods Dedicated to a Tele-Echography Robotic System", IEEE International Conference on Engineering in Medicine and Biology Society (EMBS), Istambul 2001.
38. B. Migeon, **C. Rosenberger**, P. Marché : "Correction of zoomed morphology-based interpolation of contours", IEEE International Conference on Engineering in Medicine and Biology Society (EMBS), Istambul 2001.
39. **C. Rosenberger**, K. Chehdi : "Unsupervised clustering method with optimal estimation of the number of clusters : Application to image segmentation", IEEE International Conference on Pattern Recognition (ICPR), pp. 1656-1659, vol. 1, Barcelona, Septembre 2000.
40. **C. Rosenberger**, K. Chehdi : "A New Way To Evaluate Segmentation Results", Internationnal Conference on Image and Graphics (ICIG), Tapei, Août 2000.
41. **C. Rosenberger**, C. Cariou, K. Chehdi : "Rotation Invariant Texture Analysis : A Comparative Study", EUSIPCO, Tampere, Septembre 2000.
42. **C. Rosenberger**, K. Chehdi : "Adaptive Segmentation System", IEEE International Conference on Signal Processing (ICSP), Beijing, pp. 918-921, (2), Août 2000.
43. R. Cozien, **C. Rosenberger**, P. Eyherabide, J. Rossetini, A. Ceyrolle : "Target detection and identification using neural networks and agents systems", IEEE International Conference on Information Fusion (FUSION), vol. 1, pp. MoD1-10-MoD10-17, Paris, France, July 10-13, 2000.
44. **C. Rosenberger**, K. Chehdi : "Genetic Fusion : Application to multi-components image segmentation", IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), vol. 4, pp. 2219-2222, Istanbul, Juin 2000.
45. R. Cozien, **C. Rosenberger**, P. Eyherabide, J. Rossetini, A. Ceyrolle : "Multi-agent systems and neural networks of a distributed architecture for target identification of air images", SPIE International Conference, Vol. 4118, p. 164-174, Parallel and Distributed Methods for Image Processing IV, Hongchi Shi ; Patrick C. Coffield ; Divyendu Sinha ; Eds., octobre, 2000.
46. R. Cozien, **C. Rosenberger**, P. Eyherabide, J. Rossetini, A. Ceyrolle : "Multi-agent systems and neural networks for automatic target recognition on air images" SPIE International Conference AEROSENSE, Vol. 4050, p. 139-148, Automatic Target Recognition X, Firooz A. Sadjadi Ed., août 2000.

47. **C. Rosenberger**, K. Chehdi : "Analyse Globale et Locale d'une Image", International Symposium on Image/Video Communications over fixed and Mobile Networks (ISIVC), Rabat, vol. 2, pp. 17-24, Avril 2000.
48. K. Chehdi, **C. Rosenberger**, C. Kermad : "Segmentation d'images CASI par fusion génétique", International Symposium on Image/Video Communications over fixed and Mobile Networks (ISIVC), Rabat, vol. 2, pp. 168-175, Avril 2000.
49. C. Kermad, K. Chehdi, **C. Rosenberger** : "Segmentation d'images multi-bandes par multiseuillage scalaire. Application aux images CASI", International Symposium on Image/Video Communications over fixed and Mobile Networks (ISIVC), Rabat, vol. 2, pp. 9-16, Avril 2000.
50. **C. Rosenberger**, K. Chehdi, C. Cariou, J.M. Ogier : "Texture analysis of an image by using a rotation-invariant model", IEEE International Conference on Accoustics Speech and Signal Processing (ICASSP), Phoenix, Juin 1999.
51. C. Cariou, O. Alata, **C. Rosenberger**, J.M. Ogier, K. Chehdi : "Recognition of Rotated and Scaled Textures Using 2-D AR Modeling and the Fourier-Mellin Transform", EUSIPCO, Rhodes, Septembre 1998.
52. **C. Rosenberger**, J.-M. Ogier, C. Cariou, K. Chehdi : "A statistical solution to evaluate image processing techniques", International Conference on Information Processing and Management of Uncertainty in Knowledge based Systems (IPMU), Paris, Juillet 1998.
53. J.-M. Ogier, **C. Rosenberger**, C. Cariou, K. Chehdi : "CAATI : A system based dynamic interpretation device", International Conference on Information Systems Analysis and Synthesis (ISAS), Orlando, Juillet 1998.

Conférences nationales avec comité de lecture et avec actes

1. C. Delgorge, **C. Rosenberger**, G. Poisson, P. Vieyres : "Evaluation de la qualité de résultats de compression d'images ultrasonores", Brest, CNRIUT, 2006.
2. S. Chabrier, **C. Rosenberger**, B. Emile : "Evaluation de la performance de la segmentation d'images par fusion de critères", congrès jeunes chercheurs en Vision par ordinateur ORASIS, Clermont Ferrand, France, 2005.
3. S. Chabrier, H. Laurent, **C. Rosenberger**, A. Rakotomamonjy : "Fusion de critères pour l'évaluation de résultats de segmentation d'images", 20ème Colloque GRETSI, 2005.
4. A. Chokusriwong, Romain Lavertu, Loïc Maze, **C. Rosenberger** : "Reconnaissance d'objets dans une image", Journée Pôle Capteur, Bourges, 2004.
5. H. Laurent, B. Emile, **C. Rosenberger** : "Etude comparative de critères d'évaluation de la segmentation", 19ème Colloque GRETSI, Paris, septembre 2003.
6. F. Poupeau, **C. Rosenberger**, H. Laurent, B. Emile : "Calibrage et Contrôle qualité de cerises par vision artificielle", Journée Pôle Capteur, Bourges, 2002.

7. C. Cariou, **C. Rosenberger** : "Analyse de textures invariante par rotation : Une étude comparative", Traitements et analyse d'images, Méthodes et applications (TAIMA), Hammamet, pp. 189-194, 2001.
8. C. Cassier, **C. Rosenberger**, J.-P. Louboutin, P. Marché : "Comparaison de différentes méthodes pour la segmentation et l'analyse de documents", Traitements et analyse d'images, Méthodes et applications (TAIMA), Hammamet, pp. 177-182, 2001.
9. **C. Rosenberger** : "Fusion génétique de résultats de segmentation", 18ème Colloque GRETSI, Toulouse, septembre 2001.
10. K. Chehdi, **C. Rosenberger** : "Analyse et Segmentation de textures à partir d'attributs invariants par rotation", 17ème Colloque GRETSI, Vannes, septembre 1999.

Communications sans comité de lecture et sans actes

1. C. Delgorge, **C. Rosenberger**, G. Poisson, P. Vieyres : "Evaluation de la qualité de la compression d'images ultrasonores". *GDR PRC-ISIS, juin 2006.*
2. **C. Rosenberger** : "Méthodologies d'évaluation de traitements d'images". *GDR PRC-ISIS, décembre 2005.*
3. **C. Rosenberger** : "Reconstruction 3D d'objets par une représentation fonctionnelle". *Séminaire du laboratoire d'Informatique fondamentale d'Orléans (LIFO), Orléans, 2005.*
4. **C. Rosenberger** : "Une image peut en cacher une autre". *Conférence les Mardis de la science, Bourges, 2000.*
5. **C. Rosenberger**, K. Chehdi : "Fusion génétique : Application aux images multi-composantes". *GDR PRC-ISIS, 2000.*
6. **C. Rosenberger**, K. Chehdi : "Evaluation d'un résultat de segmentation sans connaissance de la vérité de terrain". *GDR PRC-ISIS, 1999.*

Rapports de recherche

1. **C. Rosenberger** : "Rapport d'expertise ANVAR (confidentiel)" 2006.
2. "Rapport d'activités de l'équipe Signal, Image et Vision 2005-2006".
3. "Rapport d'activités de l'équipe Signal, Image et Vision 2003-2004".
4. **C. Rosenberger** : "Rapport final PTR : Contrôle qualité de cerises par vision", 2003.
5. C. Kermad, **C. Rosenberger**, M.P. Vandecandelaere, B. Vozel, K. Chehdi : "Segmentation d'images multi-composantes par fusion génétique : application aux images CASI". Rapport scientifique GSTB, février 2000.
6. **C. Rosenberger**, K. Chehdi : "Rapport Région final", *Rapport de recherche - LASTI, septembre 1999.*

7. **C. Rosenberger**, K. Chehdi : "Rapport Région 2ème année", *Rapport de recherche - LASTI, septembre 1998*.
8. **C. Rosenberger**, K. Chehdi : "Rapport Région 1ère année", *Rapport de recherche - LASTI, septembre 1997*.

PARTIE II

Résumé des activités de recherche sur l’évaluation d’algorithmes de traitement d’images

Chapitre 2

Problématique de l'évaluation d'algorithmes de traitement d'images

Ce chapitre positionne la problématique de l'évaluation pour qualifier un algorithme de traitement d'images. L'évaluation de performances est particulièrement détaillée et une classification des différents critères d'évaluation est proposée. Cette étude bibliographique permet de positionner les travaux réalisés sur l'évaluation de la segmentation et de la compression qui sont présentés dans les deux prochains chapitres.

2.1 Introduction

Le traitement d'images a suscité une multitude de travaux sur des tâches très variées : compression, tatouage, segmentation, interprétation, détection de mouvements, reconstruction 3D... [Pavlovic *et al.* 1997, Cryer et Shah 1999, Lorigo et Govindaraju 2006, Bebis et Miller 2006]. Nous considérons ici un algorithme de traitement d'images comme une boîte noire qui, pour une image originale, fournit un résultat pouvant prendre différentes formes (une image, une liste d'éléments, des valeurs numériques...). On distingue trois caractéristiques essentielles d'un algorithme (voir figure FIG. 2.1). La première correspond à l'ensemble des paramètres de l'algorithme (seuils, taille de voisinage...). Ces paramètres peuvent être fixés arbitrairement par l'opérateur humain (ou par défaut) ou par calcul en fonction de l'image originale à traiter. Le second élément correspond à la base de connaissances exploitée par l'algorithme. Il se peut que l'algorithme n'utilise pas de connaissance supplémentaire à celles utilisées lors de la conception de l'algorithme. Cette base de connaissances peut contenir des données d'apprentissage pour le traitement à proprement dit ou des règles pour fixer les paramètres de l'algorithme en fonction du contexte. Enfin, l'implémentation de l'algorithme (langage de programmation, système d'exploitation...) est aussi l'une des composantes importantes.

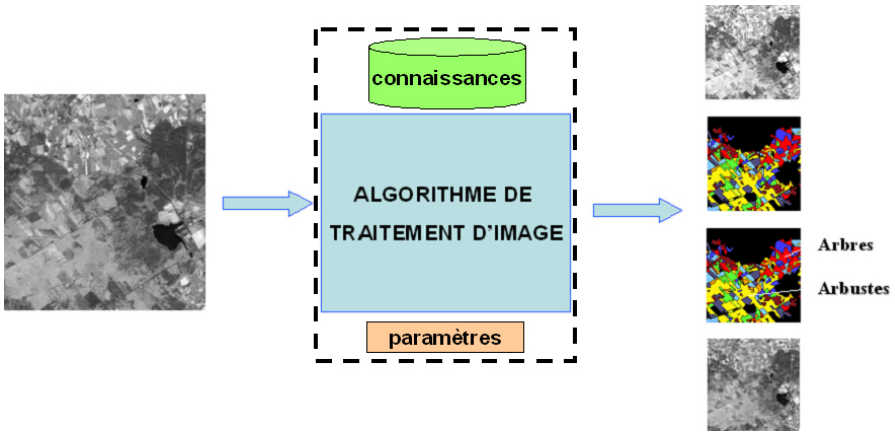


FIG. 2.1 – Un algorithme de traitement d'images

Un problème reste récurrent : comment évaluer la qualité du résultat d'un algorithme de traitement d'images ? Ce problème est primordial pour plusieurs raisons. Il est en effet absolument nécessaire pour le concepteur d'un algorithme de traitement d'images d'avoir un retour d'expérience sur son travail. Mon algorithme fonctionne t-il ? Il n'est pas toujours simple de répondre à cette question. Il est par exemple assez difficile de percevoir l'impact visuel d'une restauration d'une image (même s'il existe des tests psychovisuels).

La seconde question se posant est la suivante : quelles sont les performances intrinsèques de mon algorithme ? Ces performances doivent être en adéquation avec un cahier des charges définissant les objectifs de l'algorithme développé. Des contraintes peuvent être associées aux objectifs du traitement comme la minimisation d'un temps de traitement ou l'utilisation d'un matériel spécifique. Afin de l'appliquer dans le monde industriel, il est absolument nécessaire de qualifier la qualité du traitement d'un algorithme de traitement d'images [Forstner 1996]. Le paramétrage de l'algorithme va influencer ses performances. L'évaluation permet d'avoir un retour d'expérience sur ce paramétrage et éventuellement de l'automatiser en fonction de l'image à traiter afin de rendre le traitement adaptatif.

Dans le domaine de la recherche, une autre question se pose naturellement : dans quelle mesure mon algorithme est-il plus efficace qu'un autre de la littérature ? La possibilité d'évaluer et de comparer différents algorithmes de traitement d'images permet de choisir le plus efficace pour une tâche donnée dans une chaîne de traitement d'images. Ceci permet aussi d'identifier une technique en émergence à exploiter pour la définition de nouvelles méthodes. A titre d'exemple, les séparateurs à vaste marge ont montré leur efficacité en reconnaissance de formes et cette approche est donc devenue la méthode d'apprentissage la plus utilisée en reconnaissance d'objets dans une image [Everingham *et al.* 2006]. Dernier point, l'évaluation est la clef de voûte de la recherche

scientifique car elle permet de mettre en évidence l'apport d'une nouvelle technique au vu de la littérature. C'est donc une garantie sur l'innovation d'un nouvel algorithme.

L'évaluation d'un algorithme peut se concevoir à différents niveaux :
- qualitatif : l'algorithme me fournit-il bien le résultat escompté ?
- quantitatif : quelle est la qualité du résultat obtenu ?
- fonctionnel : quel temps dois-je attendre pour obtenir le résultat ?
- robustesse : l'algorithme rend-il toujours un résultat correct si l'information à extraire dans l'image est altérée (bruit, contraste,...) ?

Dans ce chapitre, nous nous intéressons surtout à ce dernier niveau sur la qualité du résultat. Son objectif est de recenser les différentes méthodologies dans la littérature qui permettent de quantifier la qualité d'un résultat de traitement d'images. Le plan de ce chapitre est le suivant. La section 2.2 développe les différentes méthodologies d'évaluation utilisées par la communauté scientifique. La section 2.3 présente les grandes catégories de critères d'évaluation en détaillant leurs avantages et inconvénients. L'évaluation des performances est particulièrement détaillée car elle joue un rôle essentiel dans le jugement réalisé. Nous donnons dans la section 2.4, les conclusions sur l'évaluation et les approches à privilégier.

2.2 Les méthodologies d'évaluation

C'est véritablement en 1986 [Thacker *et al.* 2005] que la communauté en vision et traitement d'images a identifié la nécessité de se doter de méthodologies et d'outils pour l'évaluation rigoureuse d'algorithmes [Haralick 1986, Price 1986]. De nombreuses organisations ou conférences ont depuis vu le jour afin d'encourager les chercheurs à développer de nouvelles méthodologies ou métriques pour l'évaluation (ECVnet[1], PETS[2]...) ou à quantifier la qualité des différents traitements de la littérature.

Nous distinguons trois grandes démarches d'évaluation d'un algorithme de traitement d'images :
- l'évaluation des performances ;
- l'évaluation par adéquation ;
- l'évaluation par diagnostic.

Ces méthodologies sont complémentaires et peuvent être utilisées conjointement. Par la suite, nous détaillons chacune d'entre elles.

[1] http ://www-prima.imag.fr/ECVNet/
[2] http ://www.cvg.rdg.ac.uk/VS/

2.2.1 Évaluation des performances

L'évaluation des performances permet un jugement qualitatif et quantitatif sur un algorithme de traitement d'images. Une propriété d'un algorithme est appréciée par un critère caractéristique, comme par exemple le PSNR, le taux de classification correcte ou un jugement subjectif.

La pertinence du critère d'évaluation utilisé est importante car elle influence de façon conséquente celle du jugement réalisé. Dans la mesure où l'évaluation des performances est souvent prépondérante, nous détaillerons dans la section 2.3 les différentes catégories de critères d'évaluation couramment utilisés en traitement d'images.

2.2.2 Évaluation par adéquation

Cette approche est classiquement utilisée en ingénierie de conception d'un système. Un algorithme de traitement d'images sera donc considéré comme étant de bonne qualité s'il remplit les besoins du cahier des charges. Cette approche permet un jugement qualitatif sur l'algorithme d'un point de vue fonctionnel. Il est tout de même nécessaire en général d'exploiter des critères quantitatifs afin d'effectuer un jugement sur la qualité d'un algorithme. Prenons en exemple le cas de la compression d'images. L'objectif principal et implicite de ce traitement, en dehors d'une application bien précise, est de réduire la taille d'une image en stockage (voir figure FIG. 2.2). Des contraintes sont généralement ajoutées comme une qualité souhaitée à un débit donné ou la minimisation des temps de compression et décompression. Afin de valider ou d'invalider l'exigence fonctionnelle d'un nouvel algorithme sur le temps de codage par exemple, il est nécessaire de calculer un critère quantitatif.

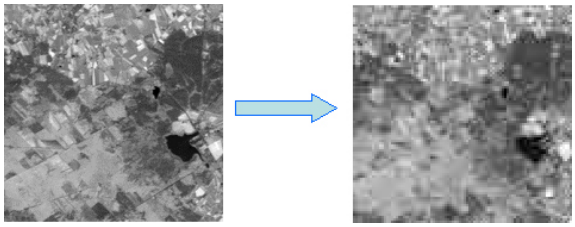


FIG. 2.2 – L'exemple de la compression d'images (compression JPEG avec un taux de compression de 20%) : à gauche, l'image originale et à droite le résultat de compression.

Cette approche n'est pas la plus utilisée dans la littérature car elle ne permet qu'une validation fonctionnelle de l'algorithme. Elle est souvent implicite. Toutefois, pour des applications spécifiques, c'est une démarche classique [Silvestre *et al.* 2004, Feng et Fulton 2005].

Deux éléments importants sont à considérer pour cette approche. D'une part, les métriques et critères de performance ont un impact important dans le jugement fonctionnel de l'algorithme. D'autre part, le jugement qualitatif d'une propriété attendue d'un algorithme n'est pas toujours simple à obtenir. En effet, si pour le cas de la compression, il est assez aisé de vérifier le gain de stockage obtenu après compression, il n'en est pas de même tout le temps. Il est parfois assez difficile de définir l'objectif d'un algorithme ou une étape comme par exemple la segmentation. Nous reviendrons plus en détail sur ce problème dans le chapitre 4 de ce manuscrit.

2.2.3 Évaluation par diagnostic

L'évaluation par diagnostic consiste à apprécier le comportement d'un algorithme de traitement d'images sur une série de tests. Un test est défini par trois éléments (voir figure FIG. 2.3) :

- la base d'images : elle contient un ensemble d'images correspondant à différents scénarii sur lesquels on veut tester l'algorithme. En général, les images sont de deux types. Le premier correspond à des images représentatives de l'application ciblée (en mode opérationnel). Le second type correspond à des cas particuliers permettant de mettre en évidence l'efficacité d'un algorithme dans un contexte difficile. Afin d'avoir un jugement pertinent sur la qualité du résultat d'un algorithme, la base d'images de test doit être évidemment la plus importante possible.
- la référence : elle doit permettre d'estimer dans quelle mesure les résultats obtenus par un algorithme sont proches de ceux attendus. Une référence correspond généralement à des vérités terrains déterminées manuellement par un expert. Il est possible aussi d'utiliser des données synthétiques.
- les métriques ou critères de performance : ils caractérisent les performances de chaque algorithme en terme de traitement (rapidité, complexité...) et de résultat (similarité du résultat obtenu avec la référence).

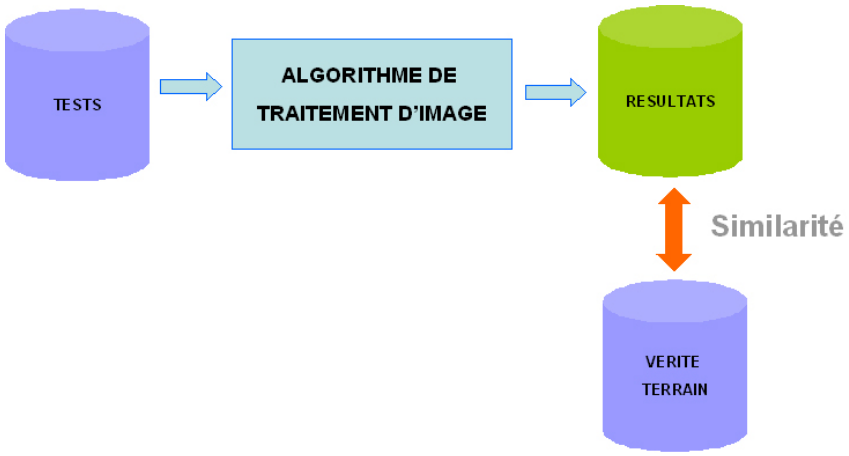


FIG. 2.3 – Evaluation par diagnostic

Différents programmes de recherche ou conférences adoptent cette méthodologie pour identifier l'algorithme le plus efficace de l'état de l'art pour une tâche donnée. Nous citons quelques exemples :

- Text Retrieval Conference[3] (TREC) : créée en 1992 et supportée par le département de la défense des États-Unis, cette conférence internationale a pour but d'évaluer les techniques d'extraction d'informations dans du texte (indexation, SPAM...). Un des objectifs est aussi d'améliorer la pertinence des méthodologies d'évaluation de ce type d'algorithmes. Cette conférence est l'un des précurseurs dans l'évaluation de technologies algorithmiques ;
- TREC Video Retrieval Evaluation[4] (TREC VIDEO) : cette conférence était une session de la conférence TREC en 2001 et 2002. Depuis 2003, son objectif est d'évaluer les techniques d'extraction d'informations de vidéos. Elle est organisée comme une compétition et met à disposition des participants des bases de données vidéo, des métriques d'évaluation pour les différentes tâches (détection d'un individu dans la vidéo, segmentation en plans...) et des vérités terrain générées par des experts ;
- IEEE International Workshop on Performance Evaluation of Tracking and Surveillance[5] (PETS) : conférence internationale créée en 2000 sur le thème spécifique de la vidéosurveillance et de son évaluation ;
- PASCAL : ce réseau d'excellence européen organise notamment une compétition appelée VOC[6] (de l'anglais "Visual Object Classes") sur la détection et la reconnaissance d'objets en mettant à disposition des participants des vérités-terrain, une base d'images et des métriques. La figure FIG. 2.4 présente deux exemples de vérité-terrain utilisée dans le cadre de cette compétition (contour de l'objet extrait manuellement, boîte englobante et région associée à l'objet) ;

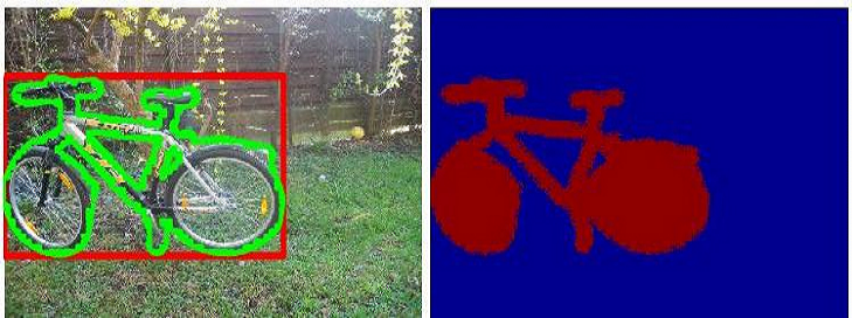


FIG. 2.4 – Deux exemples de vérité terrain utilisée dans la compétition VOC ici sur la détection d'un vélo dans une image

[3] http ://trec.nist.gov/
[4] http ://www-nlpir.nist.gov/projects/trecvid/
[5] http ://www.cvg.rdg.ac.uk/VS/
[6] http ://www.pascal-network.org/challenges/VOC/

- TECHNOLANGUE[7] : cette action du gouvernement français lancée en 2001 a pour but de faciliter une veille technologique et l'évaluation des outils de traitement de la langue française ;
- TECHNOVISION[8] : cette nouvelle action du gouvernement français lancée en 2004 vise la création d'une dynamique de l'évaluation de technologies de vision par ordinateur. Différents projets ont été retenus et procéderont, en adoptant l'approche d'évaluation par diagnostic, à la comparaison d'algorithmes de la littérature dans différents domaines (ROBIN[9] : détection et reconnaissance d'objets, EPEIRES[10] : traitements de documents...).

Afin d'avoir un jugement le plus pertinent possible, la base de tests doit être très importante. La référence associée est par conséquent assez fastidieuse à obtenir. La qualité de la référence (justesse, précision...) est aussi primordiale car elle conditionne grandement la qualité de l'évaluation d'un algorithme. Il n'est pas rare que les résultats issus d'algorithmes lors de ces compétitions permettent de mettre en évidence une erreur ou une imprécision dans la vérité terrain extraite manuellement. Enfin, les métriques utilisées pour comparer les résultats attendus et ceux obtenus par un algorithme ont une influence non négligeable sur la pertinence de l'évaluation. Même si cette approche est sans doute la plus complète, elle nécessite beaucoup de rigueur dans la définition de ces 3 éléments, ce qui en fait une approche très coûteuse.

Après cette présentation des trois grandes méthodologies, nous détaillons de façon plus précise les différents types de critères d'évaluation utilisés dans la littérature.

2.3 Les critères d'évaluation

Un critère d'évaluation permet de quantifier dans quelle mesure une propriété est vérifiée par un algorithme de traitement d'images. Certains critères permettent un jugement absolu du résultat d'un traitement d'une image, comme par exemple, le taux de compression d'une image. Dans ce cas, un critère quantifie une caractéristique d'un algorithme. Dans certains cas, il est plus facile d'accéder à un jugement relatif sur le traitement d'une image. La valeur du critère d'évaluation permet alors de comparer deux résultats issus de deux algorithmes différents ou du même algorithme avec des paramétrages différents. Cette information est moins riche mais en général plus facilement accessible.

Nous distinguons dans la littérature trois catégories de critères d'évaluation que nous allons expliciter par la suite :

[7] http ://www.technolangue.net/
[8] http ://www.recherche.gouv.fr/appel/2004/technovision.htm
[9] http ://robin.inrialpes.fr/
[10] http ://epeires.loria.fr/tiki/tiki-index.php ?page=Projet+Epeires

– les critères supervisés ;
– les critères non supervisés ;
– critères hybrides.

2.3.1 Les critères supervisés

Le principe des critères supervisés est de tirer parti d'une connaissance supplémentaire pour effectuer le jugement sur la qualité d'un résultat issu d'un algorithme de traitement d'images (voir figure FIG. 2.5). Cette connaissance est en général une référence représentant un résultat idéal ou un jugement considéré comme pertinent. Cette référence peut prendre plusieurs formes, entraînant un jugement plus ou moins automatisable.

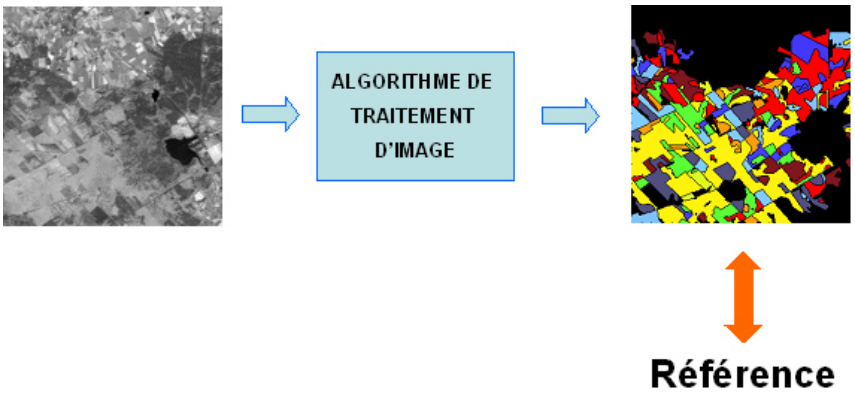


FIG. 2.5 – Evaluation supervisée

2.3.1.1 Les données synthétiques

Il est possible d'apprécier la qualité d'un traitement en l'appliquant sur des images synthétiques. Dans ce cas, il n'y a aucune incertitude liée à la référence. L'évaluation est réalisée en mesurant la similarité entre le résultat obtenu par un algorithme et cette référence. L'évaluation peut être automatisée grâce à l'utilisation de cette métrique de similarité ce qui permet d'utiliser une base d'images très conséquente pour l'évaluation.

La figure FIG. 2.6 présente quelques exemples d'images synthétiques. Si l'on prend l'exemple du filtrage d'une image, l'altération d'une image peut être modélisée (ajout de bruit) et appliquée sur une image. Le résultat d'un algorithme de filtrage sera considérée de bonne qualité s'il maximise sa similarité avec l'image originale [Aiazzi *et al.* 1997]. Dans le cas de la stéréovision, il est possible de générer deux images correspondant à deux caméras virtuelles dans un environnement 3D pour confronter le résultat de reconstruction 3D par stéréovision avec les données synthétiques [Bougeniere *et al.* 2005].

L'inconvénient majeur de cette approche est que les images ne sont pas complètement représentatives des images à traiter en mode opérationnel. Il est certes possible d'essayer de reproduire les conditions d'acquisition, le bruit du capteur et autres artefacts [Argenti et Alparone 2002].

FIG. 2.6 – Exemples d'images synthétiques

2.3.1.2 Le jugement subjectif

Ce type de critère subjectif est le plus naturel et consiste à prendre en compte le jugement d'un ou plusieurs experts. Encore une fois, cette prise en compte d'une expertise peut être réalisée de différentes façons. D'un point de vue historique, l'une des premières approches a consisté à utiliser des images dites "benchmark" comme l'image bien connue "Lena" (voir figure FIG. 2.7). Cette image constitue en quelque sorte un cas d'école. L'évaluation d'un algorithme sera estimée à partir des quelques images censées représenter un contexte général d'utilisation. Le résultat optimal d'un traitement est supposé connu par tous ou tout au moins, plusieurs résultats sur cette image par différents algorithmes sont accessibles. L'évaluateur humain doit prendre en compte ces deux types d'information (image et résultats de l'état de l'art) pour définir son jugement sur la qualité d'un nouveau résultat. S'il est possible d'apprécier (dans une certaine mesure) la qualité globale d'un traitement, il reste très difficile d'estimer un apport d'un algorithme par rapport à ceux de l'état de l'art. Malgré tout, cette approche d'évaluation a été très largement utilisée dans les dernières décennies pour justifier de la qualité d'un nouvel algorithme. A l'heure actuelle, cette approche est utilisée de façon plus marginale ou à titre indicatif, ce qui est une bonne chose.

Une autre solution consiste à demander à des experts de générer manuellement le résultat optimal ou escompté du traitement (quand c'est possible). C'est une démarche couramment utilisée en détection d'objets dans une image par exemple. En général, on essaie de fusionner les vérités terrain générées par différents experts afin d'obtenir une référence représentative. Afin d'évaluer la qualité du résultat généré par un algorithme, une mesure de similarité avec la référence experte est utilisée. Cette approche a l'avantage d'avoir une relation étroite entre l'évaluation et l'application souhaitée. La figure FIG. 2.8 présente quelques vérités-terrain extraites manuellement de segmentation d'images

FIG. 2.7 – Exemple d'une image benchmark

naturelles [Martin *et al.* 2001]. Un des inconvénients de cette approche est le coût de génération de la référence (ici obtenue par différents étudiants de l'université de Berkeley). La fusion des vérités terrain provenant de différents experts est aussi une tâche délicate, notamment concernant la gestion des informations complémentaires. En effet, si l'on observe les résultats générés par les experts, le niveau de précision des résultats de segmentation est assez disparate. Ceci montre qu'il faut spécifier de façon très précise la tâche demandée aux experts.

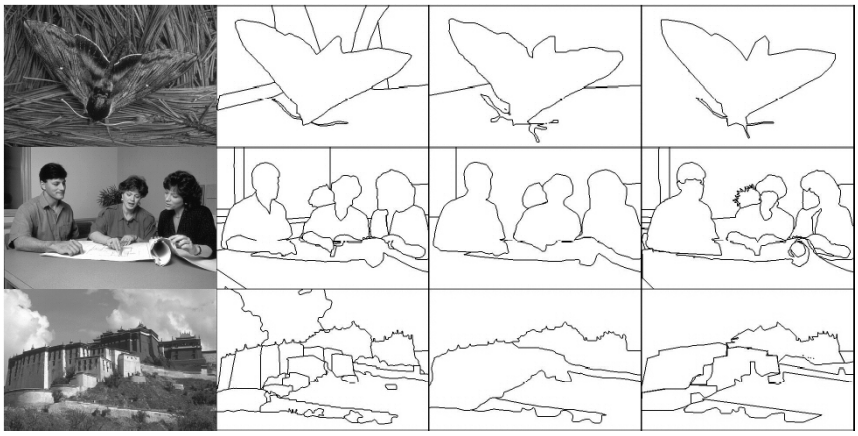


FIG. 2.8 – Exemples de vérités-terrain réalisées manuellement pour la segmentation d'images naturelles par trois experts.

Certains traitements d'une image ont pour but la génération d'une image avec une qualité proche de l'image originale, pour améliorer sa qualité (filtrage, restauration...), pour y dissimuler des informations en essayant de conserver la même qualité (tatouage) ou enfin pour diminuer la taille de stockage en ayant une qualité d'image la moins altérée possible (compression). L'appréciation de la qualité subjective perçue par un observateur peut servir de référence en matière de qualité d'image [Klein 1993] pour ce type de traitement. Le Comité Consultatif International de la Radiocommunication (CCIR) a défini des méthodes normalisées pour l'évaluation de la qualité d'une image. La base

d'un test subjectif repose sur une structure normalisée, qui est constituée d'un groupe d'observateurs, d'une méthode de traduction de l'appréciation de la qualité visuelle (c'est à dire une échelle subjective de notation de la qualité), d'un ensemble d'images test, d'un matériel de visionnage, d'une salle de test et d'un protocole définissant le déroulement de la séance de notation des images ainsi que les conditions expérimentales. Les recommandations ITU-R-BT.500-11 [ITU-R-BT.500-11 2002] définissent le protocole d'évaluation et de dépouillement des résultats. Ces études conduisent à des résultats pertinents au prix d'un protocole extrêmement contraignant. Dans la littérature, peu d'évaluations d'algorithmes respectent scrupuleusement toutes ces recommandations [Beegan *et al.* 2002].

Dans tous les cas, l'évaluation subjective peut être réalisée en demandant à des experts de comparer des résultats issus de différents traitements d'une même image. Cette évaluation relative de résultats est plus intuitive pour des experts et beaucoup moins onéreuse à mettre en place. La figure FIG. 2.9 illustre la tâche demandée à des experts médicaux de comparaison de résultats de compression d'images ultrasonores [Delgorge *et al.* 2005d]. Dans ce cas, il s'agit de classer les différents résultats de compression avec un score allant de 1 (résultat de plus mauvaise qualité) à 5 (résultat de meilleure qualité). L'image originale est présente pour que l'expert puisse percevoir sa similarité avec un résultat de compression.

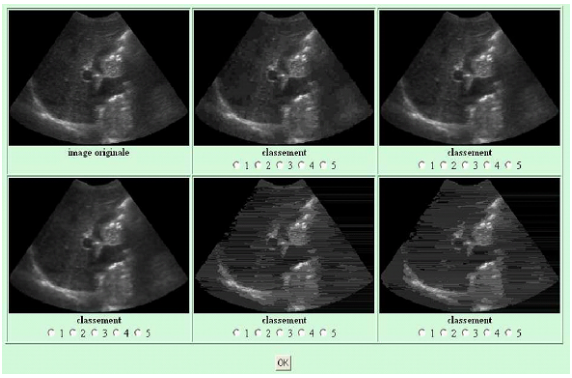


FIG. 2.9 – Comparaison de résultats de compression d'images ultrasonores

2.3.2 Les critères non supervisés

Le principe des critères non supervisés est de quantifier la qualité du résultat d'un traitement d'images à partir d'un calcul statistique (voir figure FIG. 2.10). Seule l'image originale peut être utilisée pour effectuer ce calcul. Ce type de critère mesure en quelque sorte la cohérence globale du résultat du traitement par rapport à l'image originale.

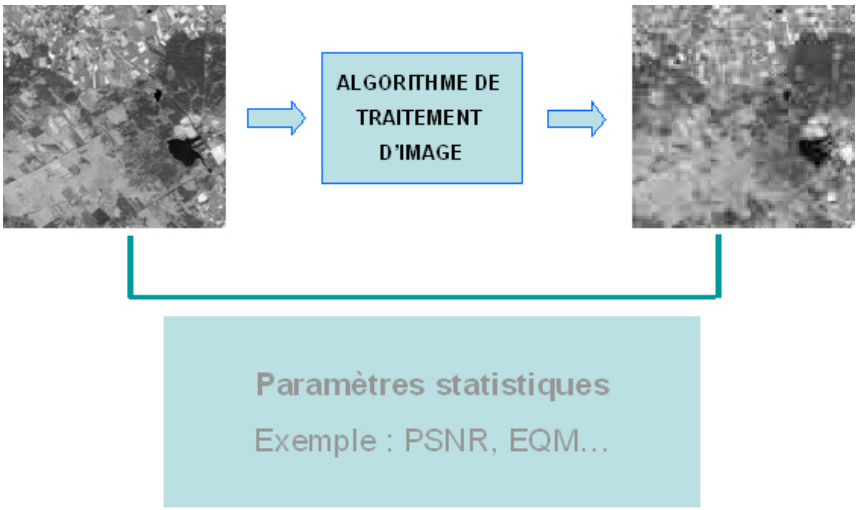


FIG. 2.10 – Evaluation non supervisée

Un des avantages d'un critère non supervisé est qu'il est en général basé sur un calcul simple. Par exemple, le PSNR largement utilisé pour évaluer un résultat de compression d'images est très simple à calculer. Le calcul de ce type de critère permet une automatisation de l'évaluation sur une grande base de résultats de traitement d'images. L'inconvénient majeur réside dans la pertinence faible ou méconnue de l'évaluation par rapport à un jugement expert. A titre d'exemple, le rapport signal à bruit de crête, ou PSNR, est le critère le plus couramment utilisé en compression [Yang 2005, Zhong et Ning 2005] même si de nombreuses études montrent que ce critère est peu pertinent [Li *et al.* 2002, Hao *et al.* 2003].

2.3.3 Les critères hybrides

L'objectif des critères hybrides est de reproduire une évaluation supervisée (considérée comme idéale) par un critère non supervisé. Il y a plusieurs stratégies pour atteindre cet objectif. La première consiste à paramétrer un critère non supervisé à partir d'informations supplémentaires ou subjectives. Il est possible par exemple, d'ajuster d'éventuels paramètres du critère non supervisé [Belaroussi *et al.* 2002] ou d'en combiner plusieurs afin d'améliorer leur pertinence [Chabrier *et al.* 2005b]. Un apprentissage peut être aussi réalisé afin d'avoir un jugement proche de l'expert à partir d'un ou plusieurs critères calculés sur le résultat de traitement d'images [Delgorge *et al.* 2005c].

Une autre stratégie consiste à intégrer dans le calcul du critère, des éléments du jugement de l'expert. Ainsi, en qualité subjective d'images par exemple, on intègre des notions de perception humaine dans le calcul du critère (critères basés sur le système visuel humain) [Callet et Barba 2001, Havstad *et al.* 2004].

Ce type de critère tire parti des avantages des deux catégories précédentes. Ils sont en général plus pertinents mais difficiles à concevoir.

2.4 Conclusion

L'évaluation est à l'heure actuelle un domaine en plein essor témoignant d'une forte maturité du traitement d'images. L'évaluation par diagnostic est incontestablement l'approche la plus utilisée par l'ensemble des organisations, groupes de recherche ou conférences. Le coût inhérent à ce type d'approche est partagé par l'ensemble des acteurs et des outils ont été développés afin de limiter celui de la constitution des vérités-terrain (logiciels d'annotation, images synthétiques réalistes...). Cette approche bénéficie grandement de l'essor du "World Wide Web" dans la mesure où il est plus facile désormais de partager des ressources notamment pour l'évaluation d'algorithmes de traitement d'images.

L'évaluation d'un algorithme de traitement d'images nécessite une bonne formalisation de ses objectifs. Le tableau TAB. 2.4 présente les avantages et inconvénients de chaque approche explicitée précédemment. Les méthodes les plus pertinentes sont celles qui font intervenir des experts. Les méthodes d'évaluation les plus simples à mettre en oeuvre consistent à utiliser un critère statistique ou une image benchmark au détriment de la pertinence. L'objectif des chercheurs en évaluation des performances est de reproduire au mieux le jugement d'expert par un critère statistique. Ceci permet une automatisation de l'évaluation en adoptant notamment l'approche d'évaluation par diagnostic.

Approche	**Méthode**	**Pertinence**	**Simplicité de mise en oeuvre**	**Rapidité de l'évaluation**
évaluation supervisée				
	vérité terrain synthétique	+++	++	+++
	image benchmark	+	++++	++++
	vérité terrain experte	++++	+	+++
	test subjectif	++++	+	+
évaluation non supervisée		++	++++	+++
évaluation hybride		+++	++++	+++

TAB. 2.1 – Tableau récapitulatif des méthodes d'évaluation de performance

Quelles que soient les méthodologies explicitées, se dégage l'importance de l'évaluation des performances. Les critères d'évaluation sont en effet à la base de tout jugement qualitatif ou quantitatif d'un algorithme. Se pose alors le problème de la validation de ces critères de performance. Dans la mesure où un critère permet un jugement sur un algorithme de traitement d'images, il doit être le plus pertinent possible. Comment quantifier

la pertinence d'un critère d'évaluation ? Les méthodologies présentées précédemment doivent être appliquées et si possible avec une rigueur accrue (base de tests conséquente, analyse de la cohérence des résultats...). C'est une condition nécessaire pour que ce critère d'évaluation soit utilisé par la communauté scientifique. La validation d'un critère objectif est souvent réalisée à l'aide de critères subjectifs. A titre d'exemple, le groupe d'experts de la qualité vidéo (VQEG[11]) adopte cette approche pour comparer l'évaluation par des critères non supervisés ou hybrides avec des critères subjectifs impliquant le jugement d'experts. Des mesures de corrélation (Pearson, Spearman) sont utilisées pour estimer la similarité de comportement.

Les deux chapitres suivants présentent des travaux réalisés sur l'évaluation de la compression et la segmentation d'images. L'objectif est d'une part d'étudier la pertinence des méthodes d'évaluation pour chacun de ces traitements. D'autre part, de nouveaux critères ont été développés afin d'augmenter la confiance dans l'évaluation.

[11] http ://www.vqeg.org/

Chapitre 3

Evaluation de la compression d'images ultrasonores

Ce chapitre relate des travaux réalisés dans le cadre de la thèse de Cécile Delgorge dont j'étais l'un des co-encadrants. Ma contribution à l'encadrement de cette thèse a principalement porté sur l'un de ses objectifs, à savoir la définition de nouveaux outils d'évaluation de résultats de compression d'images ultrasonores. Les approches présentées dans le chapitre précédent sont appliquées à ce traitement particulier et sont comparées. De nouveaux critères d'évaluation basés sur une fusion sont proposés permettant de mieux reproduire le jugement de l'expert qui est ici un médecin.

3.1 Contexte

Domaine en plein essor à notre époque, la télémédecine est née de la rencontre de la médecine et des télécommunications. Depuis 1995, le laboratoire Vision et Robotique (LVR) travaille sur le concept de la télé-échographie robotisée conjointement avec le Professeur Arbeille du service d'échographie et de médecine nucléaire de l'hôpital Trousseau de Tours. L'objectif initial de cette collaboration était de concevoir un système robotisé permettant un examen échographique de spationautes depuis la terre par un expert médical.

Le concept de la télé-échographie robotisée permet à un médecin de réaliser un examen échographique sur un patient distant. Un opérateur doit positionner sur le patient un robot manipulant une sonde échographique. Ce robot reproduit les gestes réalisés par le médecin sur une sonde fictive depuis le site distant. Le médecin visualise les images ultrasonores transmises afin de faire son diagnostic. De nombreux travaux ont été réalisés d'une part, pour valider le concept dans des conditions réelles [Gourdon *et al.* 1999] et d'autre part, pour améliorer l'ergonomie du système

[Blanc *et al.* 1999, Courreges *et al.* 2001].

Le concept OTELO (mObile Tele-Echography using an ultra-Light rObot), développé par le consortium de même nom et piloté par le LVR, repose sur un besoin identique avec des contraintes supplémentaires (voir figure FIG. 3.1). L'objectif du projet OTELO (qui a débuté en 2001) est de concevoir un système de télé-échographie robotisé portable. Le premier type d'application visé concerne les interventions d'urgence. Ceci permet d'une part de limiter les risques liés au transport du patient et d'autre part, de limiter les déplacements de l'expert médical pour des raisons d'efficacité. Le système OTELO permet aussi de réaliser des examens échographiques de patients dans des sites délocalisés depuis des hôpitaux distants.

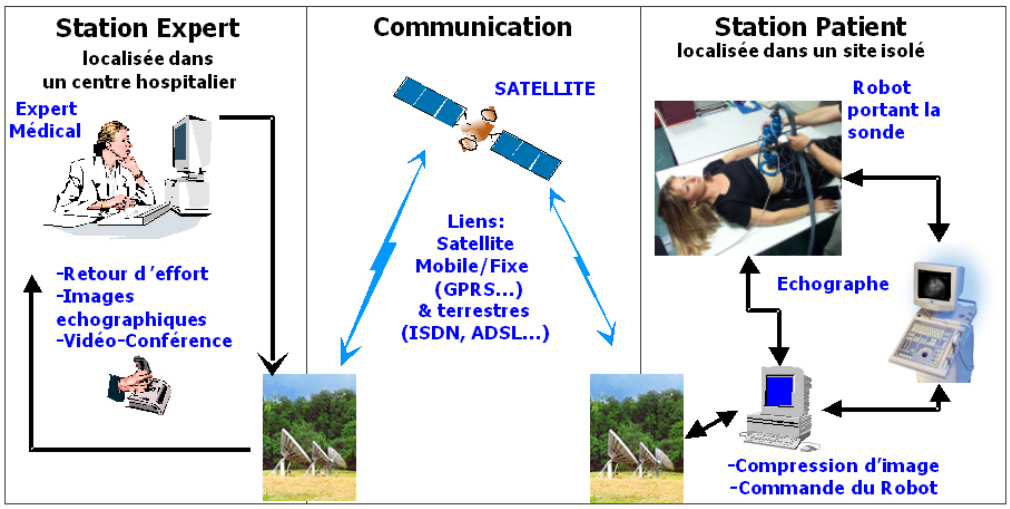
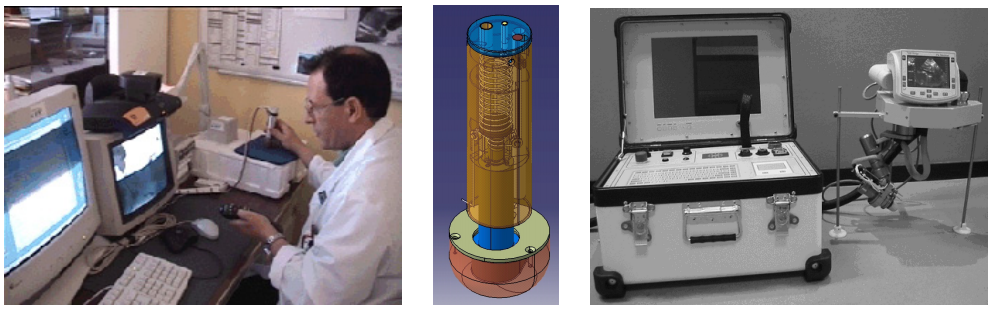


FIG. 3.1 – La chaîne de télé-échographie OTELO.

Le poste expert (voir figure FIG. 3.2 (a)) se compose d'un PC, d'une sonde fictive munie d'un capteur de localisation FOB©(voir figure FIG. 3.2 (b)). Le poste patient (voir figure FIG. 3.2 (c)) est complètement mobile, grâce au choix d'un échographe de faible encombrement, d'un robot léger, et d'une baie de commande intégrée dans une valise.

Pour pouvoir être utilisé à n'importe quel endroit, le système doit pouvoir se connecter à divers moyens de communication selon la disponibilité sur le site patient (satellite, RNIS[1], UMTS[2] ...). A l'heure actuelle, le territoire européen est encore peu couvert par le haut-débit (le taux de pénétration moyen du haut-débit dans les foyers en Europe est

[1]RNIS : Réseau Numérique à Intégration de Services
[2]UMTS : Universal Mobile Telecommunications System



(a) (b) (c)

FIG. 3.2 – Le système OTELO : (a) le poste expert, (b) la sonde fictive et (c) le poste patient.

de l'ordre de 30% d'après la source IDATE de mai 2005). Lorsque la bande passante disponible est faible, la difficulté réside dans la transmission des images échographiques qui ne doit pas engendrer une gêne trop importante pour le médecin lors de son examen.

La qualité des images reçues par l'expert doit être suffisante pour lui permettre de réaliser son diagnostic dans les meilleures conditions. La taille importante de l'image échographique engendre un long temps de transfert sur les réseaux de communication classiques (bas débit). Une étape de compression des images ultrasonores est nécessaire pour fournir, de la station patient jusqu'à la station expert des images ultrasonores dans un délai raisonnable. Si l'image reçue par le médecin ne doit subir aucune dégradation, une méthode de compression réversible doit être utilisée. Dans le cas où une perte est tolérée, une méthode irréversible est préconisée [Delgorge *et al.* 2005a].

Deux types de données échographiques peuvent être transmises à l'expert : soit des images fixes, soit une séquence vidéo. La compression vidéo est utilisée lorsque le rafraîchissement est privilégié à la qualité de l'image. Cette approche correspond à la phase de *recherche de l'organe* à diagnostiquer par l'expert médical. Tandis que la compression d'images fixes favorise la qualité au rafraîchissement. Elle a son intérêt dans la phase du *diagnostic*. Les travaux sur la compression vidéo ont été réalisés par une équipe grecque du projet OTELO [Zachariadis *et al.* 2002]. Les travaux concernant la compression des images fixes ont été réalisés par Cécile Delgorge dans le cadre de ses travaux de thèse de doctorat au sein du LVR. Un de ses objectifs de thèse était de définir une méthode d'évaluation de la qualité d'un résultat de compression d'une image ultrasonore. C'est sur ce thème que j'ai particulièrement apporté ma contribution lors du co-encadrement.

Nous présentons dans un premier temps les différentes approches d'évaluation utilisées dans la littérature afin de quantifier la qualité de méthodes de compression. Dans un

second temps, nous présentons notre contribution sur ce sujet. Les différentes perspectives de ces travaux sont présentées.

3.2 Etat de l'art

La compression d'une image, contrairement à beaucoup d'autres traitements, est un problème bien posé. En effet, l'objectif premier de cette étape est de diminuer la taille de stockage d'une image tout en limitant la distorsion par rapport à l'image originale (cette distortion est nulle en mode sans perte). De nombreuses autres caractéristiques peuvent s'ajouter comme par exemple la minimisation des temps de compression et décompression ou la robustesse vis à vis d'une erreur de transmission.

Même si ce problème est bien posé, il n'en reste pas moins difficile d'évaluer des méthodes de compression notamment en ce qui concerne la quantification de la distortion engendrée en mode irréversible (voir figure FIG. 3.3).

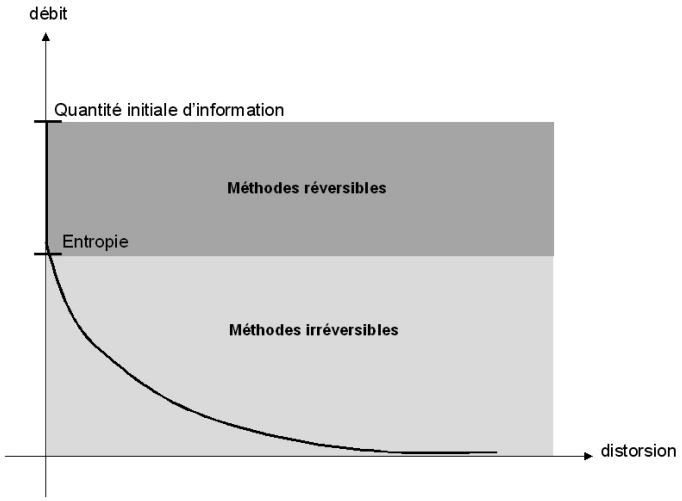


FIG. 3.3 – Mise en évidence de la distortion engendrée par la compression d'une image.

Nous présentons par la suite un bref état de l'art des méthodologies d'évaluation d'algorithmes de compression d'images.

3.2.1 Cas réversible

Dans le cas réversible, une méthode de compression est essentiellement caractérisée par le taux de compression et les temps de compression/décompression.

La taille de l'image compressée est une donnée primordiale car elle influence directement la durée de la transmission. L'efficacité de la réduction de la taille de l'image originale est souvent définie par le taux de compression C_R défini soit comme un pourcentage % :

$$C_R = \frac{y}{x}$$

soit comme un facteur :

$$C_R = (x/y) : (1)$$

où x est la taille de l'image originale et y celle de l'image après compression.

Les temps de compression et décompression représentent une mesure de performance importante pour le système OTELO. En effet, les délais peuvent perturber le contrôle du robot distant ainsi que la précision de l'acquisition de l'image échographique. Les temps de compression et décompression se cumulent au temps de transmission de l'image entre l'échographe (point d'acquisition) et l'écran du spécialiste (point de visualisation) pour constituer le délai global de transfert.

3.2.2 Cas irréversible

Dans le cas irréversible, il est nécessaire d'apprécier la distorsion engendrée par l'étape de compression. Nous distinguons trois types de méthodologie (voir figure FIG. 3.4) [Winkler 2005]. L'évaluation peut être réalisée dans un premier temps à l'aide de critères d'évaluation qui sont essentiellement des mesures quantitatives. Dans de nombreuses applications, et notamment dans le domaine médical, l'image dont on veut évaluer la qualité est destinée à être visualisée par une personne. Des méthodes d'évaluation subjectives basées sur la qualité visuelle du résultat de compression ont été définies dans un deuxième temps, impliquant le jugement d'experts. Depuis quelques années, la notion de perception humaine est prise en compte dans la définition de nouvelles mesures, les critères hybrides. Nous présentons un bref état de l'art de chacune de ces méthodologies.

3.2.2.1 Critères subjectifs

Dans le monde audiovisuel, l'appréciation de la qualité subjective perçue par un observateur demeure la seule référence en matière de qualité d'images [Klein 1993]. Le Comité Consultatif International de la Radiocommunication (CCIR) a défini des méthodes normalisées pour l'évaluation de la qualité de l'image. La base d'un test subjectif repose sur une structure normalisée, qui est constituée d'un groupe d'observateurs, d'une méthode de traduction de l'appréciation de la qualité visuelle (c'est à dire une échelle subjective de notation de la qualité), d'un ensemble d'images test, d'un matériel de visionnage, d'une salle de test et d'un protocole définissant le déroulement de la séance de notation des images ainsi que les conditions expérimentales.

Les trois tests les plus courants sont les suivants [ITU-R-BT.500-11 2002] :

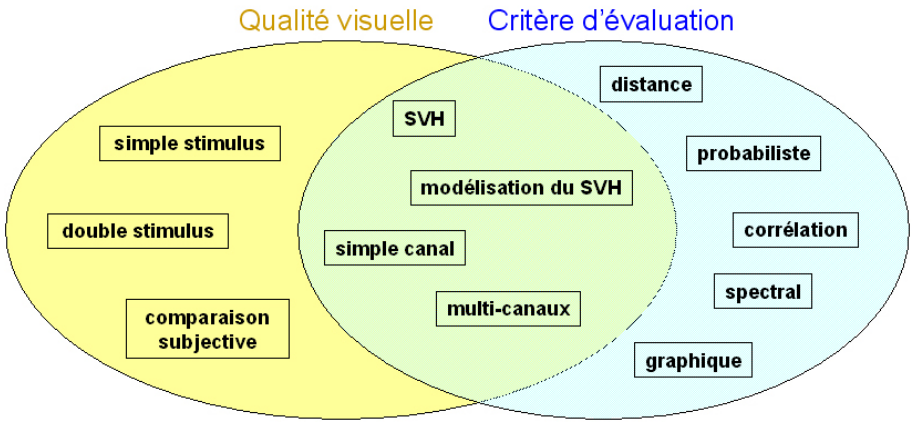


FIG. 3.4 – Différentes méthodes d'évaluation de la qualité d'un résultat de compression d'une image dans la littérature.

- **La méthode à un stimulus** : elle permet d'attribuer une qualité globale à des systèmes d'imagerie. Une série d'images dégradées par différents traitements est présentée dans un ordre aléatoire. Une note d'appréciation de la qualité est attribuée sans image de référence. L'échelle de notation peut être définie ou alors construite par l'observateur lui-même au fur et à mesure que le test avance. Ce type d'échelle libre est plus difficile à mettre en place et à analyser car il nécessite un certain nombre de séquences d'entraînement des observateurs afin de stabiliser les échelles de note.
- **La méthode à double stimulus** : les images sont évaluées par paires : une image originale de référence et une image dégradée. L'observateur doit attribuer une note à l'image dégradée selon une échelle prédéfinie. En général, deux catégories d'échelle sont proposées : la première traduit la sensation de la qualité visuelle de l'image (le critère "belle image"), la seconde traduit le degré de gêne dû à la dégradation (le critère "bonne image").
- **La méthode comparative** : Des paires d'images représentant des traitements différents sont présentées à l'observateur. La comparaison ne se fait plus par rapport à une référence mais entre couples d'images. L'observateur doit indiquer sa préférence pour l'une des deux images à évaluer. Différentes échelles peuvent être utilisées selon l'information recherchée ou le niveau de précision souhaité.

Mis à part le choix de la méthode de test, une quantité importante de recommandations est éditée [ITU-R-BT.500-11 2002]. Elles portent principalement sur :

- l'environnement du test : conditions de visionnage de la pièce, matériel utilisé, distance de l'évaluateur à l'écran,

- la résolution et le contraste de l'écran,
- la base d'images test,
- les observateurs (ou évaluateurs) : ils doivent être au moins 15, leur acuité visuelle doit être mesurée et corrigée si nécessaire, leur expérience doit être relevée (spécialité, âge),
- le protocole du test : il doit être précisément expliqué aux évaluateurs,
- la séance de test : elle doit durer au moins 30 mn,
- l'analyse et la présentation des résultats.

Ce type de test impliquant des experts suit un protocole très rigoureux et nécessite des moyens humains et financiers assez lourds. C'est pourquoi, il n'est pas rare de trouver dans la littérature des études s'inspirant de ces recommandations mais ne les suivant pas toutes [Naegele-Jackson *et al.* 2002, Azpiroz-Leehan *et al.* 2004]. De telles évaluations qualitatives et subjectives dépendent, de plus, de la spécialité médicale de l'expert, de sa motivation, de sa disponibilité, de sa fatigue ou de son entraînement à de tels tests.

3.2.2.2 Critères objectifs

La qualité d'un résultat de compression peut être représentée à l'aide d'une **distance** par rapport à l'image originale [Rosenfeld et Kak 1976]. La mesure la plus connue et la plus utilisée est celle de Minkowski (dont ses dérivées sont les distances euclidienne, de Manhattan et de Chebychev). En normalisant la distance par rapport au nombre de pixels de l'image, on obtient d'autres distances telles que la différence moyenne ou l'erreur quadratique moyenne. Ces distances peuvent être aussi calculées sur un voisinage [Tamtaoui *et al.* 1999]. La définition de ces mesures reste toutefois dépendante du type de dégradations contenues dans l'image. Il est possible de combiner plusieurs informations comme la luminance et le contraste des deux images à comparer [Wang *et al.* 2004]. Le rapport signal à bruit de crête, ou PSNR, est le critère le plus couramment utilisé [Yang 2005, Zhong et Ning 2005] même si de nombreuses études montrent que ce critère est peu pertinent [Li *et al.* 2002, Hao *et al.* 2003].

La similarité entre deux images peut se traduire à l'aide de **mesures de corrélation**. E.H. Linfoot [Linfoot 1958] a proposé différents critères se basant sur l'utilisation de la densité spectrale de puissance (DSP) des images. Le contenu structural est défini comme le rapport des DSP de l'image originale et de celle traitée. Le critère de corrélation est une corrélation des spectres normalisés par la DSP de l'image de référence. Le coefficient de Czekanowski mesure la corrélation de deux images [Androustos *et al.* 1998].

Les mesures de distance, vues auparavant, peuvent naturellement se transposer au domaine **spectral** où l'on va chercher à isoler la distorsion de l'image sur certaines bandes de fréquences, ou à distinguer la distorsion de la phase ou du module du spectre de l'image. Ces mesures ont montré leur intérêt pour l'évaluation de méthodes de com-

pression (Jpeg et SPIHT), elles sont sensibles à tout type d'artefacts [Avcibas *et al.* 2002].

T-J Chen propose dans [Chen *et al.* 2003] une nouvelle mesure de qualité utilisant les **statistiques** de Moran. Le coefficient de Moran traduit la netteté de la région étudiée, il mesure une distorsion de structure et non une variation du niveau de gris du pixel. Ce critère mesure la qualité globale de l'image mais est aussi sensible aux variations locales. Dans l'étude de Turaga [Turaga *et al.* 2004], les propriétés statistiques de l'image à évaluer sont utilisées pour évaluer la distorsion due à une compression. L'avantage est de permettre ensuite le calcul de métriques mathématiques comme le PSNR, sans disposer de l'image originale.

Des critères **graphiques** basés sur les histogrammes des images ont été proposés ces dernières années. Sundersingh [Sundersingh 2000] utilise l'histogramme de l'image différence comme mesure de fidélité entre une image ultrasonore compressée et l'image originale. La distance proposée dans [Cha et Srihari 2002] mesure la dissimilarité entre deux histogrammes.

3.2.2.3 Critères hybrides

L'objectif de ces critères est de donner des résultats d'évaluation proches de ceux obtenus lors d'une expertise subjective en conservant la simplicité de calcul des critères objectifs. Dès les années 1950, des modèles ont été proposés visant à simuler le système visuel humain (SVH), la perception visuelle. Les premiers modèles du SVH sont basés sur une approche **simple-canal**. Dans ce cas, le SVH est considéré comme un simple filtre spatial dont les caractéristiques sont définies par une fonction de sensibilité au contraste.

Le premier modèle mathématique de ce type est celui de O. Shade [Shade 1956] dont le but était de prédire la sensibilité du SVH pour la conception d'écrans de télévision. J.L. Mannos et D.J. Sakrison proposent en 1974 [Mannos et Sakrison 1974] un critère de qualité d'une image basé sur la fonction de sensibilité au contraste (FSC). Cette approche est l'une des premières mesures de qualité d'une image se basant sur un modèle de vision humaine pour l'évaluation de techniques de traitement d'images. Les deux images (l'originale et celle dont on veut évaluer la qualité) sont transformées par une fonction non linéaire d'adaptation à la luminance. Une FSC est appliquée puis la différence entre les deux images s'obtient par l'erreur quadratique moyenne. Le modèle de C.F. Hall et E.L. Hall [Hall et Hall 1977] tient compte de la non linéarité et de la réponse fréquentielle du SVH. La différence avec le modèle de Mannos et Sakrison est l'utilisation successive d'un filtre passe-bas (représentant la formation de l'image sur la rétine), d'une fonction non linéaire (représentant la sensibilité des cellules de la rétine) et d'un filtre passe-haut (représentant la formation de l'image neuronale). J.O. Limb propose un modèle [Limb 1979] en incluant un filtrage de l'erreur par un filtre passe-bas et une pondération par une fonction de masquage. La prise en compte de l'effet de

masquage est utilisée ici pour la première fois, mais les résultats sont assez médiocres. Miyahara et Algazi [Miyahara *et al.* 1996] proposent une méthodologie de mesure de la qualité d'images appelée PQS (picture quality scale) par combinaison de plusieurs critères simple canal.

Les modèles simple-canal sont toujours utilisés pour leur simplicité de calcul. Néanmoins, ils restent assez limités pour quantifier la distorsion d'une image traitée par rapport à l'image originale. La sensibilité au contraste pour un stimulus dépend de sa fréquence et de son orientation. Un modèle **multi-canaux** est un modèle de perception constitué en bandes de fréquences et d'orientation. Cette modélisation permet la prise en compte des propriétés du cortex visuel. Un modèle multi-canaux comporte un ensemble de fonctions de transfert (contrairement à une seule pour les modèles simple-canal).

L'un des modèles les plus connus de la littérature est le prédicteur de différences visuelles (de l'anglais VDP : Visual Differences Predictor) de S. Daly [Daly 1992]. Il permet d'obtenir les zones de différences perceptuelles entre une image traitée et une image originale. Ce modèle est constitué d'un ensemble de sous-modèles prenant en compte les variations de sensibilité du SVH : les variations en fonction de l'intensité lumineuse, du contenu fréquentiel et du contenu structurel de l'image. Ce critère donne une bonne représentation de la sensibilité visuelle mais demande des temps de calcul très importants. Plus récemment, Le Callet [Callet et Barba 2001] propose sur ce principe, un critère pour l'évaluation d'un résultat de compression. Il effectue la combinaison d'erreurs perceptuelles entre l'image originale et l'image compressée pour obtenir une note de qualité. Les performances du critère (évaluées à l'aide d'une évaluation subjective) permettent un gain de de l'ordre de 30% d'efficacité (mesurée par le coefficient de corrélation de rang) par rapport au PSNR. Des modèles mathématiques d'observateurs ont été développés pour la médecine nucléaire [Pommert et Höhne 2002]. L'image compressée, l'image d'erreur ou encore un critère statistique simple sont ensuite pondérés par un de ces modèles. La définition de modèles artificiels d'observateurs relève d'une procédure assez longue et coûteuse et est dépendante du type d'images testées.

Certaines études modélisent le SVH puis élaborent des critères à partir de ce modèle [Carnec 2004]. On peut trouver également dans la littérature de nombreux critères proposant de combiner une métrique statistique et certaines caractéristiques du SVH. Dans l'étude [Miyaji *et al.* 2000], le SNR est pondéré par des facteurs locaux reflétant la perception visuelle (l'effet de masque, les caractéristiques fréquentielles). Lorsque certaines spécificités de l'image sont ainsi prises en compte, on peut constater que la corrélation du critère avec la qualité subjective augmente (par rapport à l'utilisation de la métrique statistique seule). L'erreur d'évaluation constatée par ce critère diminue de l'ordre de 30% celle réalisée en utilisant le PSNR.

3.2.3 Discussion

Les critères de performance sont couramment utilisés dans le cas réversible et ne posent aucun problème pour la comparaison de méthodes de compression réversibles. Les temps de compression et de décompression restent néanmoins dépendants de l'implémentation de l'algorithme et de la machine utilisée (vitesse du processeur, mémoire vive...). Ceci est à prendre en compte lors d'une évaluation absolue d'un algorithme.

Dans la littérature, nous constatons qu'un grand nombre de mesures est proposé pour l'évaluation de la qualité de la compression dans le cas irréversible. L'état de l'art dressé ci-avant n'est certes pas exhaustif, mais représente une bonne vue d'ensemble. L'utilisation d'un critère objectif tire son avantage d'une grande simplicité de mise en oeuvre et d'analyse des résultats. Ces différents critères présentent en général une assez faible corrélation avec la notion subjective de la qualité de l'image. Les mesures sont basées sur une différence pixel à pixel, alors que l'oeil humain tient compte du voisinage. La mesure d'un tel critère statistique ne représente alors pas une mesure significative et suffisante de qualité visuelle de l'image. Malgré tout, la mesure la plus utilisée pour évaluer la qualité d'une image reste le PSNR.

Les critères hybrides proposés ne reposent pas complètement sur des résultats expérimentaux. Ces modèles sont soit trop simplistes pour réellement représenter la perception visuelle d'un humain, soit trop complexes pour représenter un critère facilement utilisable.

Afin de pallier ces différents inconvénients, divers travaux ont tenté de combiner plusieurs mesures statistiques et de les rapprocher du SVH. L'article [Avcibas *et al.* 2002] présente une étude statistique de 26 mesures objectives pour l'évaluation de méthodes de compression. Le comportement de ces mesures et leur sensibilité à certaines dégradations sont étudiées par des techniques d'analyse de la variance. Leurs similarités sont représentées sur une carte de Kohonen. Les métriques donnant les meilleurs résultats sont combinées, définissant alors une «mesure améliorée». Les mesures spectrales, la distance de multirésolution ont montré, entre autre, leur simplicité et leur meilleure sensibilité aux artefacts de codage. Burgul [Burgul *et al.* 2000] mesure la qualité d'une image échographique dans un contexte de télé-échographie. Dans une région d'intérêt sont calculés la NMSE (Normalized Mean Square Error), l'histogramme, un tracé de surface, le spectre de Fourier et des mesures de contraste. La NMSE se révèle la métrique la plus sensible à la dégradation et l'analyse de Fourier est aussi un bon indicateur de dégradation. Les travaux de thèse de J-L Olivès [Olivès 1998] se basent sur le calcul d'un ensemble de critères de qualité représentatifs de la perception visuelle. Des critères statistiques et des critères liés au SVH sont testés et comparés pour l'évaluation de différents types de dégradations (bruit, flou, faux contours, bruit structuré). Cette étude montre l'intérêt d'une combinaison de différents critères : en montrant leur complémentarité pour différentes intensités de dégradations, l'auteur peut valider le choix de cet ensemble de critères pour la construction d'une nouvelle mesure de qualité d'images.

3.3 Notre contribution

Nous pouvons constater qu'il est difficile de privilégier une mesure par rapport aux autres : les études précédentes traitent souvent d'une distorsion particulière ou d'un type d'image spécifique. Notre contribution s'est articulée autour de deux axes. Le premier a concerné la comparaison de méthodes d'évaluation existantes de la compression d'images. L'objectif était notamment de déterminer l'approche la plus appropriée pour notre application en télé-échographie [Delgorge *et al.* 2001, Delgorge *et al.* 2002]. Le second concerne la proposition de nouveaux critères afin d'obtenir un jugement plus proche des experts médicaux.

3.3.1 Comparaison de méthodologies d'évaluation

L'objectif de cette étude était la détermination de la méthode de compression d'images ultrasonores qui, à un débit donné, maximise la qualité visuelle pour un médecin. La première approche que nous avons testée est une évaluation subjective à un stimulus impliquant plusieurs experts médicaux.

3.3.1.1 Evaluation subjective à un stimulus

Ce premier test a été réalisé via une interface Web afin qu'un maximum d'experts puissent participer à cette campagne d'évaluation. L'expert visualise 60 images, une par une, sans en connaître l'origine ou le traitement appliqué. Une seule image est présentée à l'expert. L'ordre d'apparition des images (en taux de compression et méthodes de compression) est aléatoire pour éviter l'effet d'apprentissage. La figure FIG. 3.5 illustre un écran de ce test.

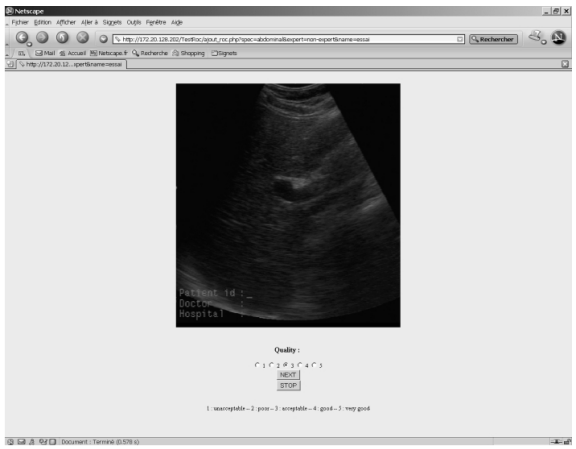


FIG. 3.5 – Evaluation subjective à un stimulus.

Il est demandé à l'examinateur d'attribuer à l'image un score entre 1 et 5, où 1 équivaut à une très faible qualité d'image et 5 à une très bonne qualité d'image. La campagne pour ce test s'est déroulée de juillet 2003 à février 2004. Ainsi, 22 tests ont été effectués au total, classés ainsi : 4 experts dont 2 évaluations complètes et 2 incomplètes, 13 non-experts et 5 tests inexploitables (abandon de l'examinateur).

Même si le test était très accessible (l'expert doit cocher sur une case pour effectuer son jugement), plusieurs difficultés ont été rencontrées et ce pour différentes raisons :

- Ce type de test à un stimulus rend nécessaire un temps d'adaptation à l'échelle des scores pour chaque examinateur. Ainsi, un référentiel se constitue pour l'expert à partir des premières images présentées (les premières réponses des experts ne sont pas fiables).
- Le test a été perçu apparemment trop long : de 30 à 45 minutes (ceci expliquant les abandons). Trop peu de tests expert ont été réalisés compte tenu de la longue durée de la campagne.
- L'examinateur n'est pas assez assisté : certains tests sont incomplets (pour cause de manque de temps, ou d'incompréhension de la méthodologie ou du protocole).

Ce test nous a permis de tirer quelques enseignements :

- Un expert préfère avoir l'image originale et les différentes résultats de traitement sur le même écran : de cette façon, l'expert possède un référentiel sur lequel il peut baser son évaluation.
- Il est nécessaire d'assister l'expert pendant le test afin d'éviter toutes incompréhensions.
- Afin d'obtenir un jugement fiable, il est nécessaire de contrôler l'environnement de l'examinateur : mesurer la luminosité de la pièce, proposer le même écran comme le stipule la norme [ITU-R-BT.500-11 2002] d'évaluation de visualisation. Cette norme est difficile à mettre en oeuvre étant données les conditions que doivent remplir l'évaluation subjective.
- La durée du test ne doit pas être trop importante pour limiter le nombre d'abandons.

L'évaluation subjective à un stimulus ne semble pas être une approche à conseiller. Elle correspond à une évaluation absolue du résultat d'un traitement qui est souvent difficile à obtenir même par un expert. Nous avons mis en place une évaluation subjective comparative en prenant en compte ce retour d'expérience.

3.3.1.2 Evaluation subjective comparative

Dans ce test, chaque écran présente une image originale comme référence et plusieurs résultats de compression (voir figure FIG. 3.6). La connaissance de l'image originale permet plus facilement d'évaluer la distorsion visuelle d'un résultat de compression [Delgorge *et al.* 2005d].

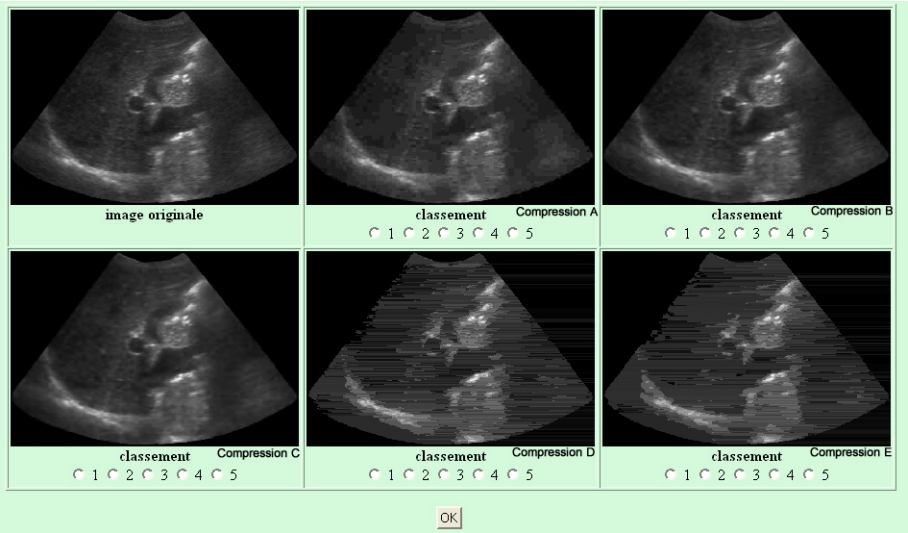


FIG. 3.6 – Evaluation subjective comparative.

Les experts doivent alors comparer et classer les résultats de compression en prenant comme référence l'image originale. Le classement "1" est attribué au résultat de compression le plus mauvais, le classement "5" est attribué au meilleur résultat. Nous obtenons par écran de test une permutation de l'ensemble $\{1, 2, 3, 4, 5\}$ (voir l'exemple donné en figure FIG. 3.7).

Les conditions expérimentales de ce test subjectif étaient les suivantes :

- il est constitué de 15 écrans comportant 5 résultats de compression d'une image originale ;
- un unique PC portable a été utilisé pour tous les examinateurs afin de garantir les mêmes conditions. Un seul écran est choisi pour toute la campagne : il s'agit d'un écran plat (moins sensible aux reflets qu'un écran classique), dont le contraste est fixé et la résolution réglée à 1024x768 à 32 bits/pixel. La résolution de l'écran vidéo est aussi fixée pour permettre de présenter le même affichage de toutes les images à chaque expert.
- la distance d'observation est identique pour tous les examinateurs.
- il n'y a pas de limitation de temps pour effectuer le test complet.
- la luminosité de la pièce est mesurée par un posemètre[3] : nous fixons la lumière incidente sur l'écran à $8,5 \pm 0,5$ et celle incidente au visage de l'examinateur à $10 \pm 0,5$.
- une assistance de l'évaluateur est assurée.

[3]Cet appareil est utilisé en photographie. Il fournit deux indications différentes sur l'ouverture du diaphragme et sur la vitesse d'obturation à adopter selon la quantité de lumière disponible dans la pièce. Il mesure en fait la lumière incidente. L'indication 8 correspond à une vitesse de $\frac{1}{125}$.

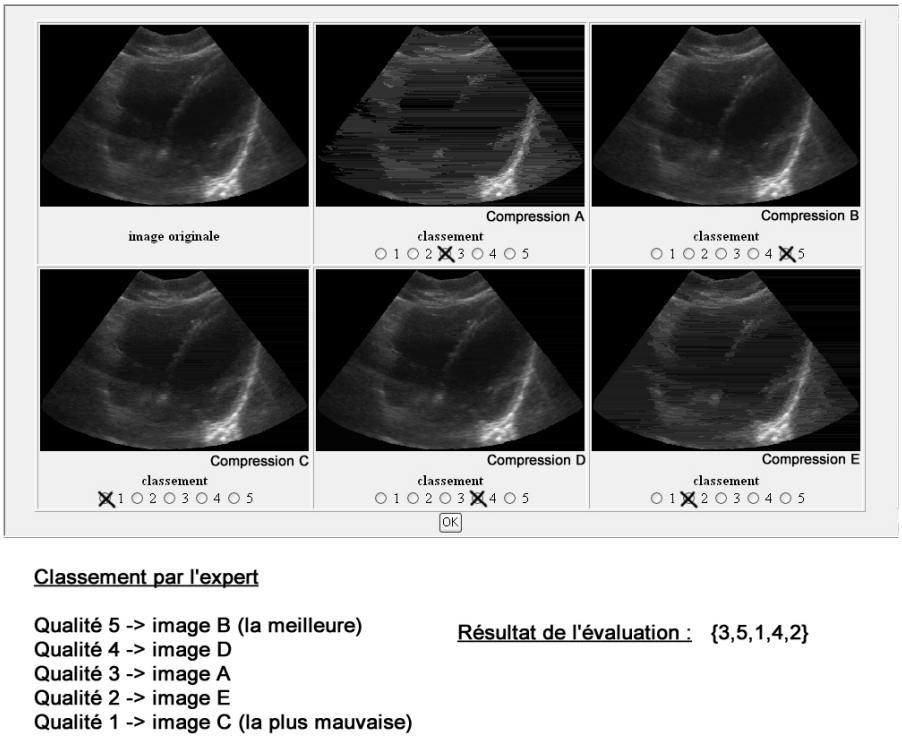


FIG. 3.7 – Classement des résultats de compression : l'image B est dans ce cas considérée comme étant la plus proche de l'image originale.

La campagne pour ce test s'est déroulée en octobre 2004. Au final, 42 tests ont été effectués au total, répartis ainsi : 13 experts médicaux en ultrasonographie (dont 10 ayant suivi le protocole) et 30 personnes non-expertes aux méthodes ultrasonores. Nous avons calculé l'écart-type des évaluations (comprises entre 1 et 5) apportées par les 10 experts ayant suivi le protocole permettant de quantifier dans quelle mesure ce jugement est un consensus. Un écart-type de 0 signifie que les 10 experts ont attribué le même classement aux 5 résultats de compression. Un écart-type de 2 (valeur maximale théorique) signifie que les 10 experts ont donné le classement le plus hétérogène à l'image. La moyenne des écarts-type des évaluations des 10 experts ayant suivi le protocole est de $0,67$. Elle est faible, ce qui montre que les réponses des examinateurs sont homogènes. Il faut noter que les écrans associés aux résultats de compression avec un débit élevé (faible compression) correspondent aux réponses les plus inhomogènes des experts, ce qui était attendu étant donnée la difficulté de comparer des résultats très proches visuellement (faible dégradation).

Notre analyse s'est basée sur l'évaluation par 10 experts : c'est un nombre important et satisfaisant, sachant que la plupart des études rencontrées dans la littérature font réfé-

rence à 3, 4 ou 5 experts [Baudin *et al.* 1996, Beaurepaire 1997, Perlmutter *et al.* 1998, Weatherburn *et al.* 2003]. Le fait d'assister l'examinateur a permis d'éviter que celui-ci interprète mal les objectifs du test. Les examinateurs se sentent plus sûrs dans leurs réponses. Leurs multiples remarques (concernant les réponses ou la structure du test) peuvent ainsi être prises en compte. Un protocole proposant une image référence s'est révélé être un bon choix : il n'y a pas de difficulté d'adaptation à l'échelle des notations ; dans l'ensemble, le test se réalise plus facilement et plus rapidement. Le protocole établi sur la luminosité autour de l'examinateur et le matériel à disposition (même écran, même PC, choix d'un écran plat moins sensible aux reflets qu'un écran classique) a permis d'homogénéiser la perception de chaque examinateur pour les images. La procédure était rapide grâce à l'utilisation du posemètre.

Cette étude s'est révélée pertinente et nous a permis de conclure sur le choix de la méthode de compression avec pertes la plus appropriée dans le cadre de notre application en télé-échographie. Néanmoins, cette étude a été longue à mettre en place. Elle a nécessité la disponibilité de beaucoup d'experts médicaux et une assistance constante. Un problème majeur de cette approche est que si nous avons d'autres résultats de compression à évaluer, il faut recommencer toute l'étude avec des experts médicaux. Afin d'éviter ceci, nous nous sommes posés la question suivante : quel critère objectif nous donnerait un jugement proche de l'expert médical ? Afin de répondre à cette question, nous avons considéré les résultats de l'étude subjective comparative par les experts médicaux comme notre référence. Ainsi, un résultat de compression A est considéré comme de meilleure qualité qu'un résultat B si le score moyen donné par les experts est supérieur pour le résultat A. Dans la section suivante, nous allons déterminer dans quelle mesure des critères objectifs de la littérature sont capables de reproduire ce jugement.

3.3.1.3 Evaluation objective

Nous avons sélectionné dans l'état de l'art précédent, 21 critères objectifs de types différents : mesures de distance, notées D_x ; mesures de corrélation C_x ; mesures spectrales S_x ; mesure de PSNR P_1 ; mesure de contraste T_1 ; mesures liées au SVH H_x. Nous avons retenu ces critères car une étude récente les avait comparées dans le cas d'images naturelles [Avcibas *et al.* 2002]. Ces critères sont présentés dans le tableau TAB. 3.1.

L'objectif de cette étude est de déterminer parmi ces 21 critères ceux qui permettent de reproduire au mieux le jugement des experts médicaux. Il est possible par exemple de calculer le facteur de corrélation linéaire de Pearson entre la valeur d'un critère et le score moyen donné par les experts sur chaque résultat de compression. Nous avons opté pour une méthode plus proche de la reconnaissance de formes. Nous avons défini une mesure estimant la similarité de comparaison de deux résultats de compression par un jugement médical et par un critère objectif. Pour chaque écran du test, nous avons un classement des 5 résultats de compression par rapport à l'image originale établi par les experts. Nous

D1	Minkowsky - Erreur moyenne absolue
D2	Minkowsky - Erreur quadratique moyenne
D3	Minkowsky - Norme infinie modifiée
D4	Erreur sur un voisinage de 8 pixels
D5	Erreur sur un voisinage de 24 pixels
D6	Erreur en multirésolution
C1	Corrélation croisée normalisée
C2	Fidélité de l'image
C3	Corrélation de Czekanowski
S1	Phase de l'erreur spectrale
S2	Phase et module de l'erreur spectrale
S3	Module de l'erreur spectrale par blocs
S4	Phase de l'erreur spectrale par blocs
S5	Phase et module de l'erreur spectrale par blocs
S6	Erreur spectrale par blocs
P1	Rapport Signal à bruit PSNR
T1	Mesure de contraste
H1	Norme SVH de l'erreur absolue normalisée
H2	Norme SVH *L2*
H3	Similarité d'indexation
H4	Erreur DCTune basée sur la transformée de la DCT

TAB. 3.1 – Les 21 critères objectifs d'évaluation de résultats de compression d'images étudiés.

pouvons déterminer le résultat de la comparaison de 2 résultats de compression parmi 5. Ainsi, du premier classement obtenu, nous pouvons déduire 10 comparaisons de couples de résultats de compression (un exemple est détaillé dans la figure FIG. 3.8). Pour chaque comparaison, la valeur 1 est donnée au résultat de compression de meilleure qualité pour le médecin et la valeur -1 à l'autre.

Un ensemble S de 150 comparaisons de résultats de compression est alors obtenu pour l'ensemble du test. Une comparaison est représentée par le couple de valeurs $\{-1, 1\}$ ou le couple $\{1, -1\}$. De cette façon, nous construisons le vecteur CE correspondant aux 150 comparaisons obtenues grâce aux classements des experts médicaux. Le vecteur CC est ensuite déterminé pour chaque critère objectif. Dans ce cas, le classement initial est établi en considérant les valeurs numériques du critère calculé sur toutes les images en prenant en compte son sens de variation. Ainsi, nous avons 21 vecteurs CC représentant chacun la comparaison des résultats de compression par un critère objectif. Pour représenter la similarité de jugement entre chaque critère objectif et notre référence donnée par

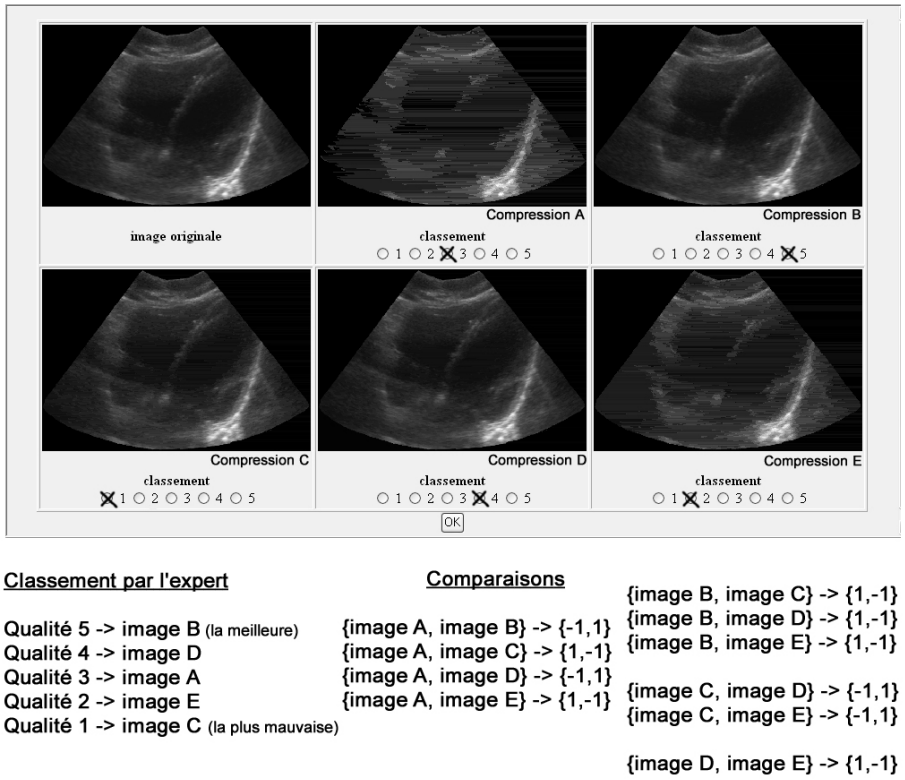


FIG. 3.8 – Principe de définition de la comparaison.

les scores moyens des experts, nous définissons un taux de bonne comparaison TBC :

$$TBC = 1 - \sum_{k=1}^{150} \frac{|CE(k) - CC(k)|}{300} \tag{3.1}$$

où $CE(k)$ et $CC(k)$ représentent respectivement le résultat de la comparaison k (de 2 résultats de compression) par les experts médicaux et par un critère objectif. Ce taux mesure la fidélité du critère à reproduire le jugement des experts (une valeur de 1 ou de 100% signifie que le critère permet d'effectuer un jugement identique à celui de l'expert). La performance d'un critère à reproduire le jugement d'un expert médical par ce taux donne des résultats similaires à ceux obtenus par le coefficient linéaire de Pearson. Néanmoins, le taux de bonne comparaison donne une information plus intuitive et plus compréhensible pour l'expert médical. Nous avons calculé la similarité TBC des comparaisons fournies par les experts et celles données par chaque critère.

La figure FIG. 3.9 présente les taux de comparaison TBC obtenues pour chacun des 21 critères. Un critère performant permet de reproduire au mieux le jugement de l'expert

doit donc obtenir une valeur de TBC élevée.

Le meilleur résultat est obtenu par D5 avec une valeur de $65,3\%$ ce qui est loin d'être satisfaisant. Ainsi, la similarité de comparaison TBC distingue 4 critères parmi les 21 : D5, T1, S2 et S1. Leurs taux de bonne comparaison sont respectivement de 65,3%, 64%, 64% et 63,3%. Nous pouvons remarquer que la mesure P1, c'est à dire le PSNR (un des critères les plus utilisés) obtient seulement la 9ème place parmi les 21 critères même s'il obtient un TBC assez peu éloigné de celui obtenu par le meilleur critère. Ce comportement non optimal confirme les résultats obtenus dans des études précédentes de l'état de l'art [Li *et al.* 2002, Hao *et al.* 2003].

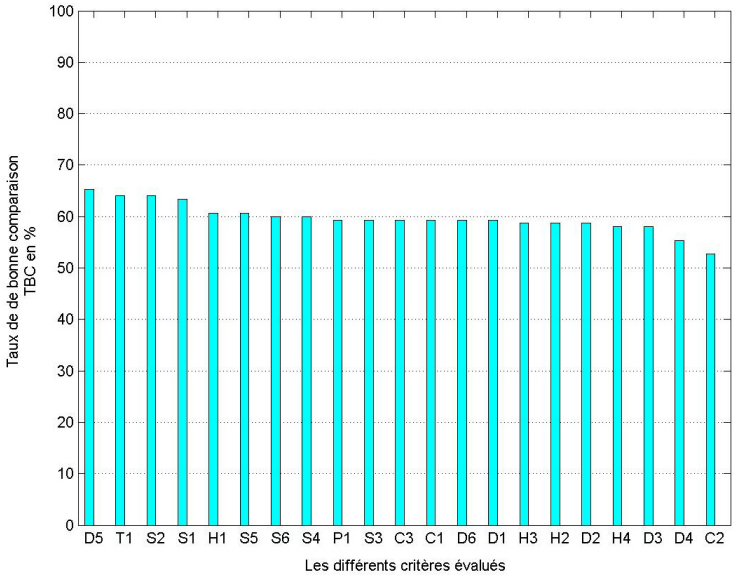
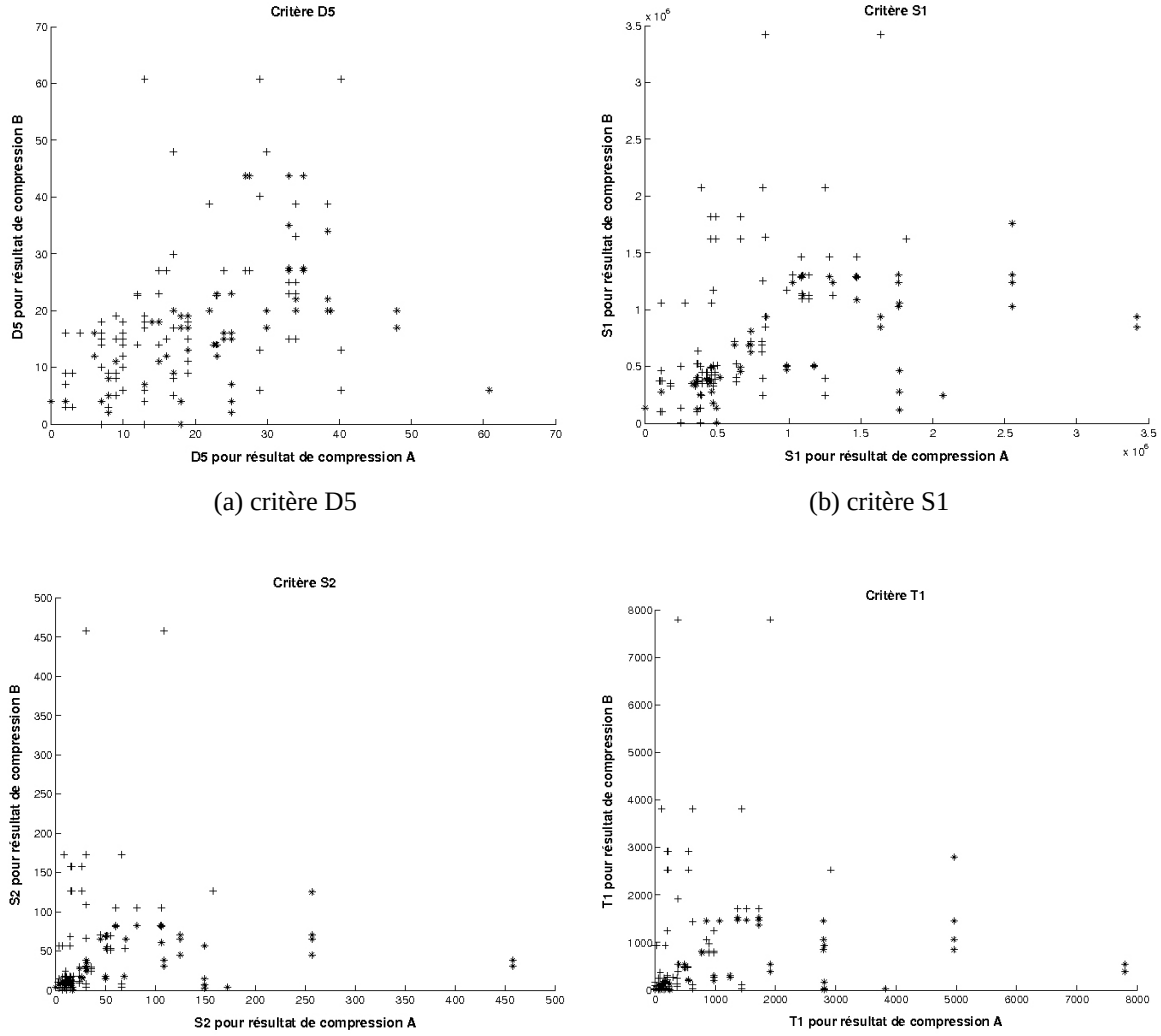


FIG. 3.9 – Taux de Bonne Comparaison TBC de chaque critère avec l'évaluation des experts (classés dans l'ordre décroissant).

Sur la figure FIG. 3.10, nous mettons en évidence la capacité des 4 meilleurs critères à reproduire le jugement expert. L'abscisse et l'ordonnée correspondent aux valeurs calculées du critère pour deux résultats de compression. Un critère sera dit pertinent si ces différents points sont facilement séparables (éloignés de la première bissectrice). On constate sur ces figures que la détermination du meilleur résultat de compression est loin d'être optimale.

La méthodologie que nous avons utilisée nous a permis d'identifier la performance des 21 critères objectifs testés. Afin d'améliorer leur efficacité, nous avons réalisé une fusion de différents critères.



(a) critère D5

(b) critère S1

(c) critère S2

(d) critère T1

FIG. 3.10 – Mise en évidence du pouvoir discriminant des critères objectifs (le symbole * signifie que le résultat de compression A est de meilleure qualité par rapport au résultat B selon les experts).

3.3.2 Fusion de critères d'évaluation

Nous souhaitons ici améliorer l'évaluation de la qualité d'un résultat de compression, et ainsi dépasser le meilleur taux de similarité obtenu, à savoir 65,3%. Divers travaux ont montré qu'il est intéressant de combiner plusieurs critères d'évaluation [Cane 1997]. On peut citer des travaux sur des combinaisons de critères par régression linéaire, par analyse de la variance et construction d'une carte de Kohonen [Avcibas *et al.* 2002] ou par algo-

rithme génétique [Olivès 1998]. Comme le montre la figure FIG. 3.11, deux approches de fusion ont été appliquées : une combinaison linéaire de critères et un apprentissage par un séparateur à vaste marge.

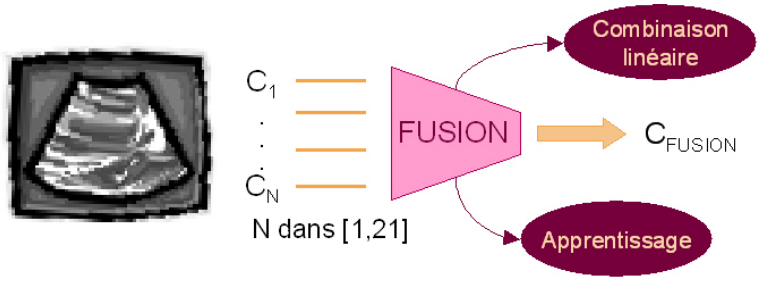


FIG. 3.11 – Fusion de critères d'évaluation objectifs.

3.3.2.1 Fusion par combinaison linéaire

La première méthode de fusion que nous avons testée consiste à combiner linéairement des critères d'évaluation [Delgorge *et al.* 2005b]. Cette approche assez simple peut être considérée comme une méthode d'apprentissage où la fonction de décision sur la comparaison de deux résultats de compression est linéaire. Dans le cas où l'on choisit de fusionner N critères, l'objectif est de déterminer les valeurs $a_i, i = 1 : N$ des coefficients de la combinaison linéaire donnant le comportement le plus proche possible du jugement clinique de l'expert. Ce premier critère noté $DEL1$ s'exprime de la façon suivante :

$$DEL1 = \sum_{i=1}^{N} a_i.C_i \tag{3.2}$$

où C_i correspond au critère objectif i. Nous avons testé deux variantes. Dans le premier cas, l'on choisit les N critères les plus efficaces d'après l'étude comparative précédente. Nous avons appelé cette approche la *combinaison linéaire sans sélection*. La deuxième approche, appelée *combinaison linéaire avec selection*, détermine les N critères objectifs à fusionner parmi les 21 ainsi que les coefficients $a_i, i = 1 : N$ associés. L'intérêt de cette dernière approche est de prendre en compte une éventuelle redondance des critères objectifs qui limiterait ainsi l'intérêt de leur fusion. Pour effectuer cette recherche de paramètres, nous utilisons un algorithme génétique. Parmi les méthodes proposées dans la littérature, nous avons choisi d'utiliser un algorithme génétique permettant conjointement la détermination des critères à fusionner et la détermination des coefficients associés. Cette méthode d'optimisation (proposée initialement par J.H. Holland [Holland 1975]) a montré son efficacité dans différentes applications

[Bhandakar *et al.* 1994, Marzouki *et al.* 1997, Yoshimura et Oe 1999].

Test préliminaire

Afin d'évaluer l'intérêt de la fusion de ces critères par combinaison linéaire, nous avons adopté un procédé classique en apprentissage. Nous avons utilisé un sous-ensemble des 150 comparaisons à notre disposition pour la détermination des paramètres et utilisé le reste pour quantifier la performance de ce critère. La figure FIG. 3.12 présente les résultats obtenus en utilisant le nombre optimal de critères pour chaque méthode. Ainsi, en utilisant 100 résultats de comparaison pour la détermination des coefficients (et des critères à utiliser pour la méthode avec sélection), nous constatons un gain dans la performance de jugement des 50 résultats de comparaisons restants par rapport à l'utilisation d'un seul critère. Ces résultats reflètent la performance moyenne obtenue sur 10 tirages aléatoires de la base d'appentissage afin de prendre en compte la complexité moyenne.

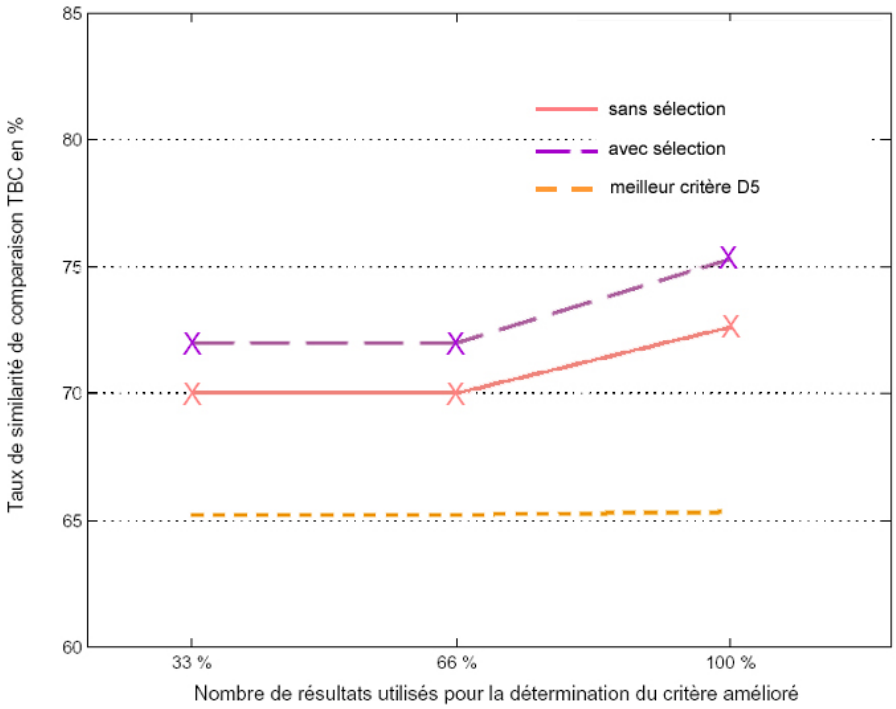


FIG. 3.12 – Généralisation de la fusion par combinaison linéaire.

Combinaison linéaire sans sélection

La figure FIG. 3.13 présente le taux de similarité de comparaison obtenu par la méthode de combinaison linéaire sans sélection. En augmentant le nombre de critères à fusionner, l'on devrait obtenir une valeur du taux de similarité croissante. En effet, en mettant le coefficient $(n+1)$ à 0 et en prenant les n coefficients trouvés précédemment, on est au moins assuré d'avoir effectivement la même fonction d'aptitude. La courbe enveloppe, tracée dans la partie supérieure de la figure, représente un minorant des valeurs que l'on aurait dû obtenir. Les valeurs constatées ici (en faisant un tirage aléatoire de chaque population initiale) sont ainsi dues à un nombre trop faible de générations écoulées ou d'individus dans la population.

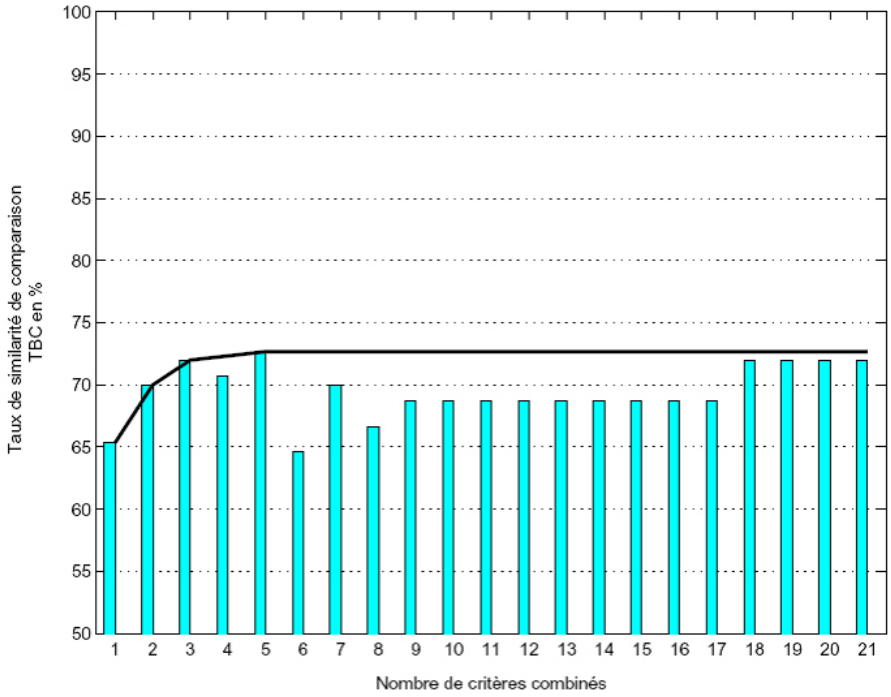


FIG. 3.13 – Taux de Bonne Comparaison TBC par fusion avec la combinaison linéaire sans sélection.

Le meilleur taux de comparaison est de $72,6\%$, il est obtenu lorsque 5 critères sont combinés. Toutes les combinaisons (sauf une) donnent une similarité de comparaison supérieure à 65%, le meilleur taux obtenu avec un seul critère était de $65,3\%$.

Afin d'effectuer une comparaison de résultats de compression plus proche de celle des experts, nous pouvons utiliser comme critère amélioré la combinaison linéaire suivante avec 5 critères :

$$DEL1 = 0,98 * D5 + 0,05 * T1 + 0,16 * S2 - 4,36.10^{-5} * S1 - 0,70 * H1 \quad (3.3)$$

Le coefficient de pondération d'un critère n'a pas réellement d'interprétation physique et ne correspond absolument pas à un coefficient d'importance du critère. En effet, les critères sont normalisés mais leurs variations peuvent être très différentes. Ces coefficients sont liés à une projection dans un espace à une dimension optimale au sens du TBC.

Nous obtenons dans ce cas $72,6\%$ de bonne comparaison sur la base de résultats testée. En présentant à l'algorithme génétique les critères dans l'ordre de meilleure performance individuelle, le résultat attendu était l'obtention d'une bonne similarité dès l'utilisation des premiers critères, puis une progression assez faible et lente. Les résultats n'affichent pas exactement ce comportement. Une explication peut être apportée du fait de la redondance inter-critères. Effectivement, après une étude portant sur la corrélation linéaire des critères, nous avons mis en évidence une forte dépendance de certains critères (S1 et S2 par exemple). Néanmoins, lorsque l'on réduit le nombre de critères à présenter à l'algorithme génétique, les similarités obtenues sont beaucoup plus faibles. Il est alors intéressant d'opérer une sélection de critères à fusionner.

Combinaison linéaire avec sélection

La figure FIG. 3.14 présente les taux de similarité de comparaison obtenus en opérant une sélection de critères. On peut remarquer que la similarité est bonne ($74,6\%$), et proche du maximum, lorsque 7 critères sont combinés. Dans ce cas, cependant, l'algorithme opère une double répétition, ce qui signifie que seulement 5 critères sont sélectionnés : $C1$ - $S2$ - $T1$ - $D5$ - $S6$ (avec présence de répétitions pour $D5$ et $S6$). Dans ce cas (et pour une similarité de $74,6\%$), le critère amélioré devient :

$$DEL2 = 0,007 * C1 - 0,19 * S2 + 0,03 * T1 + 1,05 * D5 - 0,77 * S6 \qquad (3.4)$$

La figure FIG. 3.15 montre la capacité du critère combiné à reproduire le jugement des experts. On peut remarquer que la détermination du meilleur résultat de compression est plus facile avec le critère combiné ($DEL2$). Il reste encore une confusion dans la comparaison lorsque les deux résultats de compression sont très proches (les points près de la première bissectrice ne sont pas bien séparés) ou lorsque la qualité des deux images est bonne (petite valeur du critère).

Discussion

Par cette dernière approche de combinaison linéaire avec sélection, nous améliorons la similarité de comparaison avec le jugement des experts médicaux, qui est maintenant au maximum de $75,3\%$. Nous obtenons donc un gain significatif par rapport à l'utilisation d'un unique critère (qui atteignait un TBC de $65,3\%$). Le temps nécessaire pour la détermination des coefficients et des critères à fusionner est de l'ordre de 4 minutes lorsque les calculs sont effectués sur un Pentium IV à 2,5 Ghz et pour une implémentation

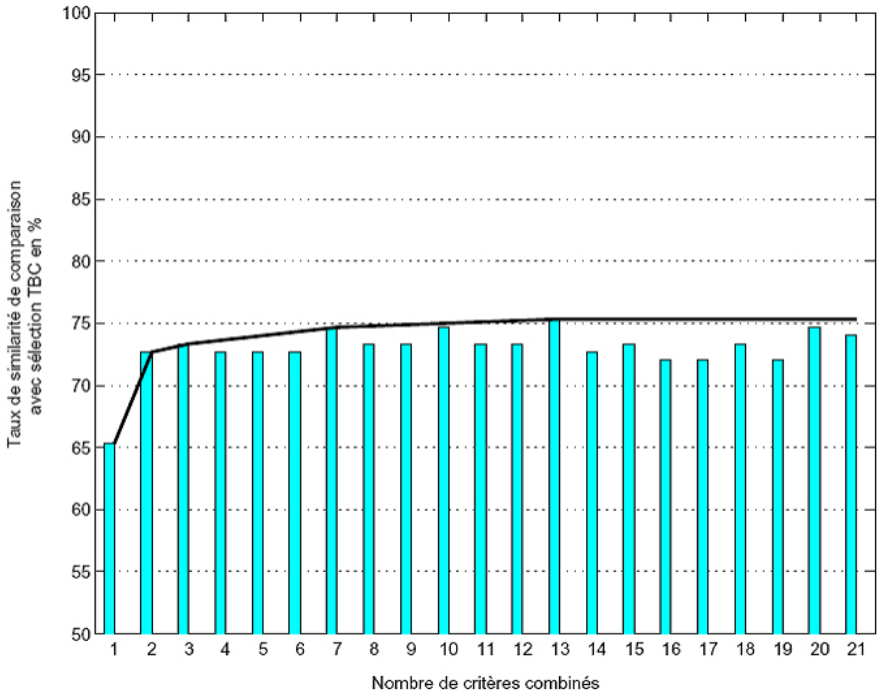


FIG. 3.14 – Taux de Bonne Comparaison TBC par fusion avec la combinaison linéaire avec sélection.

sous Matlab©[Houck *et al.* 1995a].

L'avantage de cette approche par combinaison linéaire est sa simplicité : en effet, il est nécessaire d'opérer le processus une fois pour déterminer les coefficients de la combinaison linéaire. Il suffit ensuite d'appliquer ces coefficients aux critères retenus pour trouver le nouveau critère combiné. Avec cette méthode, la décision sur la qualité relative de deux résultats de compression se fait par une fonction linéaire. Or, il n'y aucune raison que la fonction liant les évaluations des experts et les mesures des critères soit linéaire (sauf si il y avait un consensus idéal du jugement des experts). Afin de déterminer une frontière de décision plus complexe, nous avons réalisé un apprentissage par un SVM.

3.3.2.2 Fusion par apprentissage

La deuxième méthode de fusion que nous avons testée consiste à opérer un apprentissage du jugement expert médical [Delgorge *et al.* 2005c]. Les valeurs des critères calculés ainsi que l'évaluation des experts forment un ensemble d'apprentissage $\{x_i, y_i\}_{i=1,\cdot\ell}$, avec $x_i \in \mathbb{R}^{2d}$ le vecteur des $2 \times d$ critères objectifs décrivant la qualité de deux résultats de compression d'une image donnée et y_i le résultat de comparaison donné par les

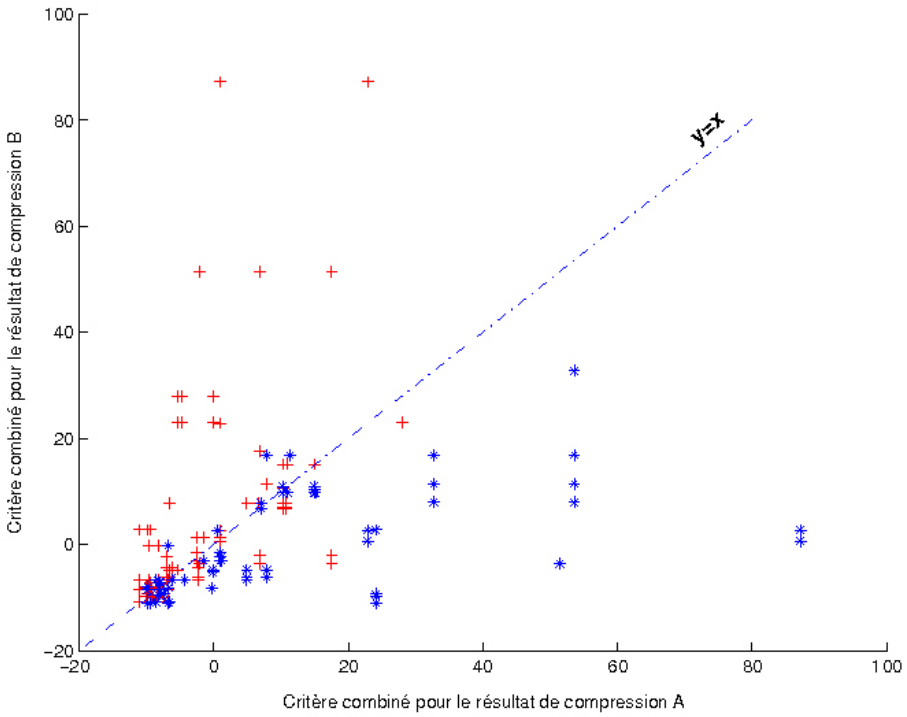


FIG. 3.15 – Mise en évidence du pouvoir discriminant du critère $DEL2$.

experts. Les évaluations des experts y_i peuvent être considérées comme des classes, et l'on se positionne dans un problème de classification. L'objectif est alors de construire une fonction f qui estime les dépendances entre les exemples de comparaison x_i et les classes y_i et qui minimise le risque d'erreur de classification, c'est à dire qui soit capable de prédire correctement le résultat de comparaison de 2 résultats de compression n'appartenant pas à la base d'apprentissage.

Généralement, dans des applications de traitement d'images, les classifieurs à base de réseaux de neurones [Singh et Amin 1999], de la méthode du plus proche voisin [Simard *et al.* 1993] ou encore des méthodes basées sur des modèles statistiques [Bailey et Srinath 1996] étaient les plus souvent utilisés. Ces méthodes ont montré leur efficacité pour la reconnaissance de caractères [Cariou *et al.* 1999] et ont été très longtemps les méthodes les plus efficaces. Les «séparateurs à vaste marge» (SVM), basés sur les concepts théoriques développés par Vapnik [Vapnik 1998], sont de plus en plus présents dans la littérature et ont prouvé leur efficacité dans des problèmes réels, tels que la reconnaissance d'images couleur [Chapelle *et al.* 1999], la reconnaissance de visages [Osuna *et al.* 1997]. C'est pourquoi, nous avons utilisé un SVM (voir figure FIG. 3.16).

La figure FIG. 3.17 présente le taux de bonne comparaison TBC obtenu selon le

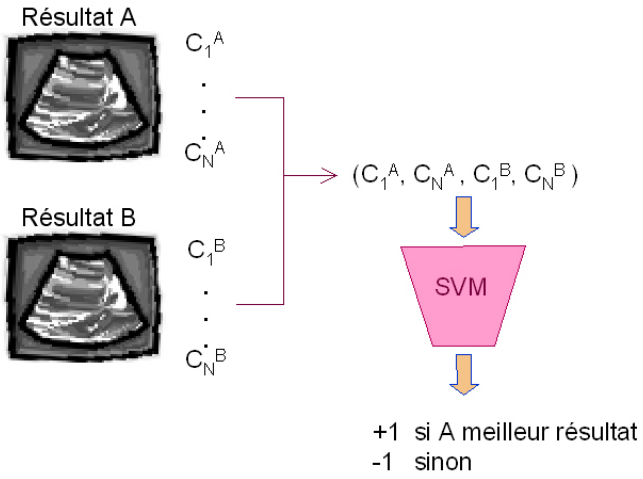


FIG. 3.16 – Fusion de critères objectifs par un SVM.

nombre de résultats de compression utilisés dans l'ensemble d'apprentissage. Lorsque 66% de la base des 150 exemples est utilisé dans l'ensemble d'apprentissage, le système obtient un taux de $86,6\%$ de bonne reconnaissance. L'apprentissage et la reconnaissance prennent en moyenne 80 secondes sur un Pentium IV à 2,5Ghz et pour l'implémentation sous Matlab©donnée dans [Canu *et al.* 2003]. Pour chaque ratio de la base d'apprentissage, nous avons lancé un grand nombre de fois l'algorithme (231 fois) dans l'objectif de s'approcher des meilleurs paramètres de l'algorithme SVM, à savoir : le paramètre C de pénalisation, le type de noyau et les paramètres liés au noyau.

Pendant l'apprentissage de cette fonction f, une sélection de variables peut être réalisée pour déterminer quelle(s) variable(s) est (sont) déterminante(s) pour la classification. L'algorithme de sélection de variables est un algorithme de classement de paramètres par retour arrière basé sur l'influence d'une variable donnée sur la marge [Rakotomamonjy 2003]. Les résultats portant sur l'étude de la sélection de variables sont montrés sur la figure FIG. 3.18 : cette étude est menée pour une base d'apprentissage de 90% des données (soit 135 résultats de comparaison) et sur les 21 critères. Le taux de bonne comparaison maximal de $92,8\%$ est obtenu pour 10 variables, c'est à dire que le nombre optimal de critères est de 5.

Un apprentissage par SVM des 150 comparaisons de résultats de compression (obtenues par comparaison des valeurs des 5 critères) nous permet d'obtenir une similarité avec le jugement des experts de l'ordre de 92.8%.

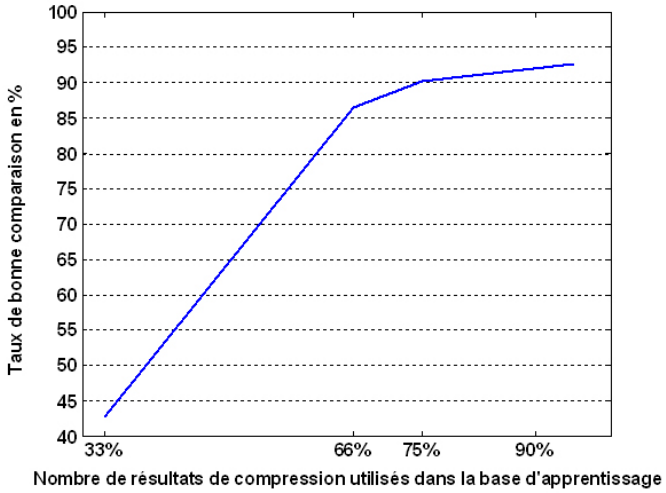


FIG. 3.17 – Généralisation de la fusion par un SVM.

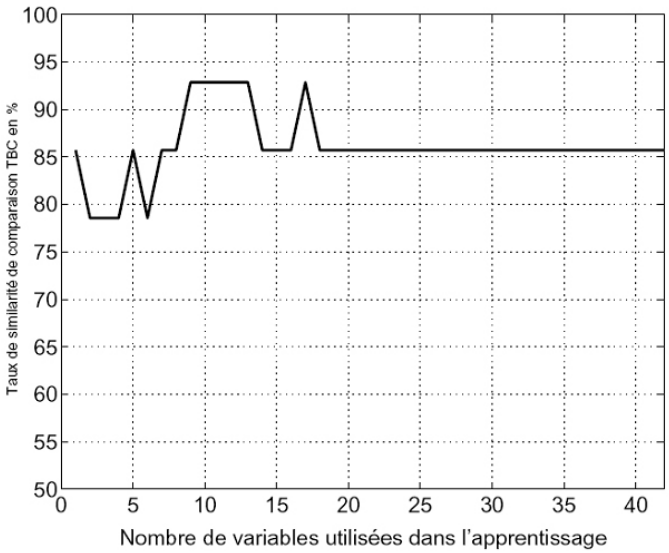


FIG. 3.18 – Sélection des variables pour la fusion avec un SVM.

3.4 Conclusion et perspectives

Notre contribution sur l'évaluation de la compression d'images ultrasonores se situe à deux niveaux.

Notre première contribution a été tout d'abord de comparer différentes méthodologies d'évaluation de méthodes de compression d'images. Le contexte de l'application médicale nous a fait naturellement choisir en premier lieu les méthodes d'évaluation subjective impliquant différents experts médicaux. Une première étude subjective à un stimulus s'est révélée difficile à exploiter. En effet, l'expert éprouvait des difficultés à ajuster son jugement car une seule image lui était présentée. Un second test prenant en compte le retour d'expérience de l'étude précédente a été réalisé. Nous avons utilisé cette fois une étude subjective comparative. Les résultats obtenus ont été pertinents, cette deuxième approche est donc selon nous à privilégier. Malgré tout, elle nécessite un grand nombre d'experts, des conditions particulières et une assistance constante. Quelques critères objectifs de la littérature ont été étudiés pour savoir dans quelle mesure ils étaient capables de reproduire le jugement médical. Le critère qui s'est révélé être le plus intéressant parmi ceux testés est basé sur l'erreur de voisinage. Néanmoins, il n'obtient qu'un taux de bonne comparaison avec l'expert de 65,3% mettant en évidence les limites des critères testés.

Notre deuxième contribution a concerné la fusion de critères objectifs d'évaluation afin de reproduire plus fidèlement le jugement d'experts médicaux. Dans ce cadre, nous avons utilisé deux approches d'apprentissage de la comparaison de résultats de compression. La première a consisté à réaliser une combinaison linéaire de critères d'évaluation pour définir un nouveau critère. Ceci correspond à spécifier le type de la fonction de décision concernant la qualité relative de deux résultats de compression. Nous avons obtenu un gain significatif de l'efficacité du jugement par la combinaison linéaire des critères testés. Par cette méthode, le taux de bonne comparaison atteint 75,3%. La deuxième méthode de fusion a consisté à utiliser un séparateur à vaste marge afin d'identifier une fonction de décision plus complexe et adaptée au jugement des experts. Cette approche a permis d'atteindre le taux de bonne comparaison de 92,8% (en utilisant comme base de test, la base d'apprentissage complète) avec le jugement des experts médicaux.

Que reste t-il à faire ? beaucoup de choses...

La comparaison des critères d'évaluation que nous avons menée n'est pas exhaustive. En effet, de nombreux critères d'évaluation de qualité d'une image sont proposés dans la littérature. Pour autant, peu de chercheurs les utilisent pour montrer l'apport d'une nouvelle méthode de compression. Quelle en est la raison ? Un processus d'évaluation ne peut être efficace que si la qualité du jugement est incontestable. Un critère d'évaluation doit être validé par un processus plus strict que les méthodes de compression que l'on va comparer. Nous avons proposé un cadre permettant de comparer différents critères objectifs. Nous avons mis à disposition de la communauté scientifique des images, des critères d'évaluation pour qu'ils soient plus souvent utilisés mais aussi pour les comparer entre eux.

La fusion n'est qu'un moyen pour augmenter la performance de l'évaluation. Il existe des multitudes de méthodes de fusion pouvant être testées afin d'améliorer encore les résultats d'évaluation. Cette façon de prendre en compte le jugement d'experts nous semble plus efficace que d'essayer d'exprimer dans un critère objectif des éléments du SVH.

Un axe de recherche intéressant est l'étude de l'influence de la connaissance *a priori* dans l'évaluation de la qualité d'un traitement d'une image médicale. En effet, la perception d'une dégradation est différente pour un expert médical ou pour une autre personne. Identifier les raisons de cette différence de jugement permettrait de définir des critères d'évaluation plus performants.

Chapitre 4

Evaluation de la segmentation d'images

L'évaluation de la segmentation d'images fait l'objet de ce chapitre. J'ai initié ces travaux lors de ma thèse de doctorat. Depuis ma nomination en tant que maître de conférences, je les ai poursuivis notamment dans le cadre du co-encadrement de la thèse de Sébastien Chabrier. Les objectifs de la segmentation sont dans un premier temps définis avant une présentation de l'état de l'art. Des travaux sous mon co-encadrement ont concerné une étude comparative des critères de la littérature et la proposition de nouveaux critères non supervisés basés sur la fusion d'un certain nombre d'entre eux. De nombreuses applications de ces critères sont ensuite présentées.

4.1 Contexte

L'objectif de ma thèse de doctorat était la conception d'un système coopératif de segmentation d'images. L'évaluation de la qualité des résultats de segmentation obtenus était une tâche incontournable afin de valider l'algorithme. Entre 1999 et 2003, j'ai essentiellement travaillé sur les applications des critères d'évaluation notamment pour la définition de nouvelles méthodes de segmentation. Ce chapitre relate aussi les activités de Sébastien Chabrier dans le cadre de sa thèse de doctorat qui a débuté en 2003 sous mon co-encadrement. L'objectif de cette thèse était d'une part de faire un recensement des critères utilisés pour l'évaluation de la segmentation d'images et d'étudier leur pertinence. D'autre part, de nouveaux critères plus performants ont été proposés.

Contrairement à la compression, la segmentation est un problème mal posé. Évaluer la qualité d'un résultat de segmentation présuppose, comme nous l'avons vu dans le chapitre 2 de ce manuscrit, de formaliser les objectifs de la segmentation. La section 4.2 de ce chapitre définit les objectifs de la segmentation. Nous présentons dans la section 4.3 un bref état de l'art des méthodes utilisées pour l'évaluation de la segmentation d'images. Nous détaillons dans la section 4.4 notre contribution au domaine et nous donnons dans

la section 4.5 les conclusions et perspectives de ces travaux.

4.2 Définitions de la segmentation

Si l'on reprend le schéma classique d'une chaîne de traitement (voir figure FIG. 4.1), la segmentation est une étape préalable à l'interprétation.

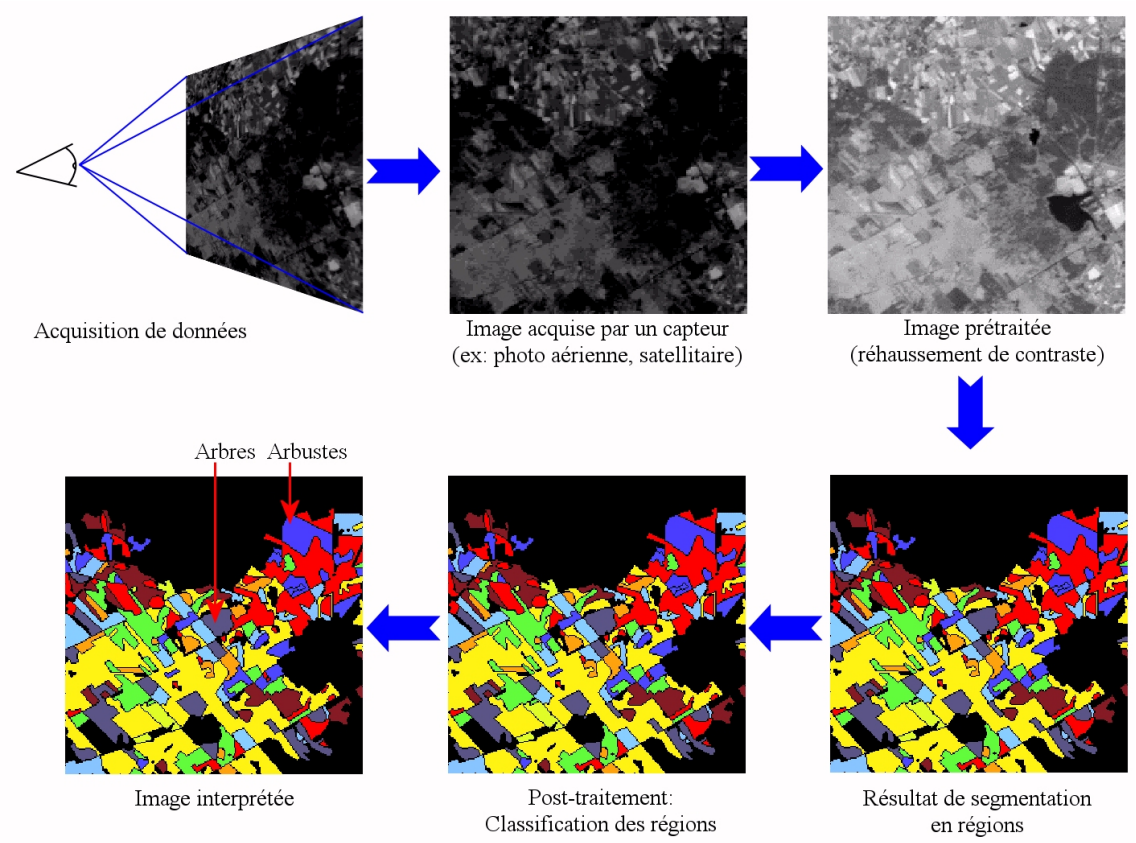


FIG. 4.1 – La chaîne de traitement d'une image (images du ©LASTI).

Un point de désaccord entre les différents chercheurs de la communauté est la définition des objectifs de la segmentation. Si nous revenons à l'origine historique de la segmentation, elle a grandement été inspirée par le système de perception visuelle humaine car l'objectif était de faciliter une interprétation automatique d'une image de manière similaire à une interprétation humaine. Le système de perception visuelle humaine utilise les notions de similarité et de différence afin de localiser les objets d'une scène ainsi que leurs frontières ou contours. Ces deux notions ont donné naissance aux deux approches couramment utilisées dans le domaine de la segmentation, à savoir respectivement les approches dites "région" et "frontière", comme le présente Cocquerez et Philipp dans [Cocquerez et Philipp 1995] : "*La segmentation fait référence aux notions*

de différence et de similarité comme les perçoit le système visuel humain et ceci donne naissance à deux approches couramment qualifiées d'approche frontière et d'approche région".

Une des plus anciennes définitions de la segmentation proposée dans la littérature et aussi une des plus utilisée, est celle d'Haralick et Shapiro [Haralick et Shapiro 1985] énonçant qu'un bon résultat de segmentation doit vérifier les quatre principes suivants :

- "Les régions d'un résultat de segmentation d'une image doivent être uniformes et homogènes par rapport à des caractéristiques comme le niveau de gris ou la texture".
- "L'intérieur des régions doit être simple et sans trop de trous".
- "Les régions adjacentes d'un résultat de segmentation doivent avoir des valeurs significativement différentes par rapport aux caractéristiques pour lesquelles elles sont uniformes".
- "Les frontières de chaque contour doivent être simples, régulières et doivent être précises spatialement".

Nous pouvons constater que ce type de définition ne nécessite aucune information sur l'application envisagée après l'étape de segmentation. Ainsi, Levine et Nazif [Levine et Nazif 85] donnent une définition similaire en précisant qu'il s'agit bien d'un procédé de bas niveau : "*Nous pouvons définir un résultat de segmentation comme une partition d'une image en régions qui possèdent un contenu uniforme et qui possèdent un contraste avec les régions voisines adjacentes. L'uniformité et le contraste sont mesurés comme un ensemble de caractéristiques de bas niveau qui peuvent être évaluées sur toute l'image*". Borsotti a une vision similaire de la segmentation d'images [Borsotti *et al.* 1998] : "*L'objectif de ces algorithmes de segmentation d'images est de partitionner de façon indépendante du domaine d'application l'image en un ensemble de régions qui sont visuellement distinctes et uniformes par rapport à des propriétés comme le niveau de gris, la texture ou la couleur*". Des définitions similaires peuvent encore être retrouvées dans divers autres articles ([Kara-Falah 1995, Beauchemin *et al.* 1998]...).

Au contraire, pour certains auteurs comme Zhang par exemple [Zhang 1997], la segmentation fait partie de l'interprétation et n'est pas une étape à part : "*La segmentation d'une image consiste à la subdiviser en parties la constituant et à extraire ces régions d'intérêt (les objets)*". Les objectifs de la segmentation sont liés à l'interprétation réalisée ultérieurement selon Andrey [Andrey 1999] : "*La segmentation d'images est un processus de bas niveau qui a pour objectif de partitionner une image en régions homogènes. La manière dont l'homogénéité des régions est définie dépend de l'application*". En ce qui concerne Mac Cane [Cane 1997], l'étape de segmentation et celle d'interprétation ne sont pas uniquement liées, mais confondues : "*En fait, les applications d'interprétation d'images sont juste des algorithmes de segmentation complexes*".

Qui a raison ? En fait, il existe selon nous, différentes catégories de méthodes de segmentation (voir figure FIG. 4.2) donc autant de définitions. La différence fondamen-

tale entre toutes ces méthodes est la quantité de connaissance *a priori* exploitée pour la segmentation d'une image. Ainsi, la frontière entre la segmentation et l'interprétation est assez difficile à définir.

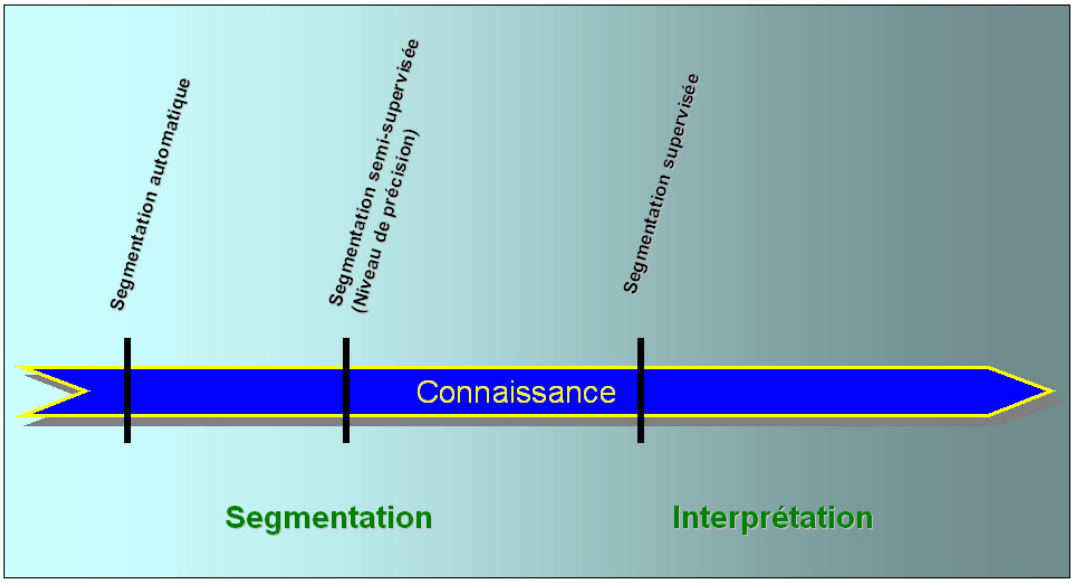


FIG. 4.2 – Les différentes catégories de méthodes de segmentation.

La première approche dite de "*segmentation automatique*" a pour objectif une partition de l'image en régions ou une détection de frontières par un procédé automatique (pouvant toutefois être réglé par des paramètres). Les objectifs de cette approche rejoignent notamment ceux donnés par Haralick et Borsotti. Il est illusoire de penser que le résultat de segmentation permettra une interprétation quelle que soit l'application envisagée. Pour autant, la segmentation automatique a un intérêt pour de nombreuses applications où l'on souhaite accéder à une partition automatique de l'image. La segmentation permet par exemple, d'obtenir un partitionnement d'une image originale pour une étape de compression [Lee et Lewicki 2002]. Elle permet aussi le calcul des descripteurs sur les régions détectées pour optimiser l'indexation [Schettini *et al.* 2002].

La deuxième dite de "*segmentation semi-supervisée*" exploite une information sur le niveau de précision du résultat de segmentation. Le niveau de précision peut être aussi défini par l'utilisateur. Ainsi, si nous prenons l'image de la figure FIG. 4.3, le niveau de précision demandé permettra de déterminer si le résultat de segmentation doit mettre en évidence chaque grain de sable ou bien une seule et même région texturée. Dans l'approche frontière, le nombre de pixels appartenant à une frontière dans le résultat de détection est généralement utilisé pour définir ce niveau de précision. Dans l'approche région, il est possible de prendre en compte le nombre de régions (ou de classes lorsque la méthode de segmentation utilisée est basée sur une classification). Nous considérons que

le nombre de classes est un indicateur plus pertinent pour définir le niveau de précision du résultat car il est indépendant du nombre d'occurrences d'un même objet dans l'image. Prenons comme exemple une photo d'une surface uniforme sur laquelle nous avons déposé des grains de riz et des grains de pois chiche comme le montrent les photos des figures FIG. 4.4 $a)$ et $d)$. On pourrait considérer que le résultat de la figure FIG. 4.4 $e)$ (38 régions) est plus précis que celui de la figure FIG. 4.4 $b)$ (7 régions) si l'on considère le nombre de régions. Pourtant, ces deux images contiennent les mêmes objets. Dans ce cas, si l'on considère le nombre de classes, nous concluons au même niveau de précision pour les deux résultats des figures FIG. 4.4 $c)$ et $f)$.

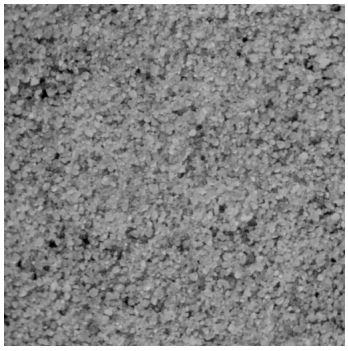


FIG. 4.3 – La texture sable.

La "*segmentation supervisée*" constitue la troisième approche de segmentation. Dans ce cas, un ensemble de connaissances est utilisé pour obtenir le résultat de segmentation. Cette approche a pour objectif une application bien particulière où par exemple, des données d'apprentissage vont être utilisées pour faciliter la détection d'objets d'intérêt dans l'image. Il y a beaucoup d'applications en imagerie médicale de ce type [Beichel *et al.* 2005, Tsai *et al.* 2005]. La question que l'on peut se poser est de savoir si cette approche fait réellement partie de l'étape de segmentation ou plutôt de celle d'interprétation. Dans la prochaine section, nous présentons une brève description des approches d'évaluation d'un résultat de segmentation dans l'état de l'art.

4.3 État de l'art

On retrouve encore une fois les deux catégories de critères d'évaluation du deuxième chapitre de ce manuscrit, à savoir l'évaluation supervisée et non supervisée.

4.3.1 Evaluation supervisée

Un résultat de segmentation est évalué en prenant en compte des connaissances *a priori*. Ces connaissances peuvent être une vérité terrain [Hoover *et al.* 1996] (segmentation de référence) ou bien des connaissances sur les éléments à identifier. Les vérités ter-

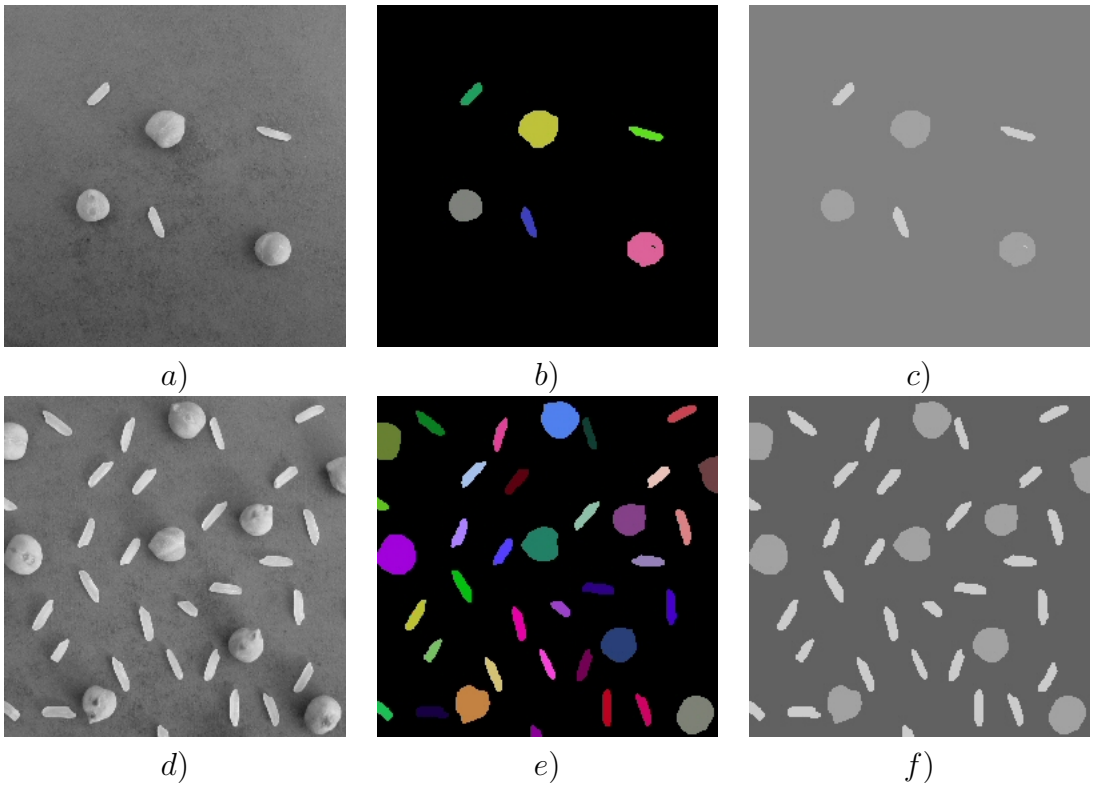


FIG. 4.4 – Différents résultats de segmentation d'une image. $a)$ image originale, $b)$ segmentation en 7 régions, $c)$ segmentation en 3 classes, $d)$ image originale, $e)$ segmentation en 38 régions, $f)$ segmentation en 3 classes.

rains peuvent être obtenues grâce à des experts ou bien en créant des images synthétiques (voir figure FIG. 4.5). Cette connaissance pour l'évaluation permet de prendre en compte de façon explicite les objectifs éventuels de la segmentation. Cette approche permet une évaluation d'un résultat pour les différentes catégories de méthodes de segmentation détaillées précédemment.

4.3.1.1 Évaluation de la détection de frontières

Dans le cas de résultats de segmentation sous la forme de frontières, l'évaluation est généralement réalisée à l'aide d'une métrique de similarité entre les contours extraits par la méthode de segmentation et ceux de la vérité terrain. L'erreur quadratique moyenne (Root Mean Square) est sans doute l'une des premières mesures de divergence proposée ([Coquin *et al.* 1997] ou exprimée dans l'espace de Fourier [Wilson *et al.* 1997]). Ces mesures de distance peuvent être complétées par différentes distances provenant de l'interprétation probabilistique d'images : les distances de Küllback et Bhattacharyya, ainsi que celle de Jensen [Basseville 1989] reposant sur les entropies de Rényi [Michel *et al.* 1994]. Ces mesures prédisent de façon très imprécise les déformations perceptibles. Comme elles n'intègrent pas l'information spatiale sur les pixels, elles peuvent

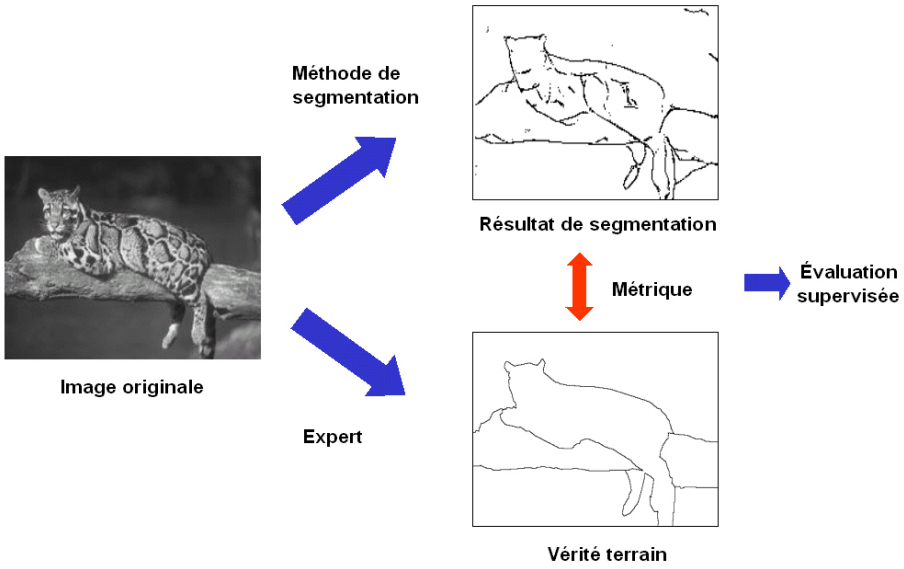


FIG. 4.5 – Évaluation supervisée d'un résultat de segmentation à l'aide d'un tracé expert.

donner lieu à des valeurs aberrantes. La distance de Hausdorff [Beauchemin *et al.* 1998] est une métrique intéressante car sans connaître aucune corrélation entre les pixels de l'image, elle donne une bonne mesure de similarité entre deux résultats de segmentation. Par contre, elle est très sensible au bruit et son temps de calcul est important. La distance de Baddeley [Baddeley 1992] est une variante de la distance de Hausdorff pour les images binaires. Elle possède aussi une grande sensibilité au bruit. Baddeley étend la mesure présentée dans [Baddeley 1992] pour les images binaires aux images en niveaux de gris dans [Wilson *et al.* 1997]. Des extensions de la distance de Baddeley ([Zamperoni et Starovoitov 1996], [Coquin *et al.* 1997], [Coquin et Bolon 2001]) sont utilisées pour mesurer la dissimilarité entre des images en niveaux de gris.

Pratt [Pratt *et al.* 1978] propose une mesure empirique de comparaison entre une carte de frontières et une carte de référence. Odet [Odet *et al.* 2002] propose une mesure de divergence échelonnable permettant d'évaluer différents niveaux d'erreur de résultats de segmentation binaires : évaluation de détection de frontières avec un faible écart par rapport à la vérité-terrain (VT) et erreur moyenne. Correira [Correia et Pereira 2000] utilise des critères spatio-temporels pondérés sur un résultat de segmentation dans une vidéo. La technique proposée par V. Barranco-Lopez [López *et al.* 1998], [Roman-Roldan *et al.* 2001] a pour objectif d'évaluer des résultats de segmentation avec une faible erreur de détection par rapport à la frontière théorique. L'expression mathématique de ce critère de qualité a été obtenue de manière simple et intuitive et possède une très bonne efficacité de calcul. Cette mesure est très souple car elle peut être adaptée à des critères plus spécifiques d'appréciation visuelle grâce à ses sept coefficients mais un entraînement préalable est nécessaire.

4.3.1.2 Évaluation de la segmentation en régions

Nous présentons ici des critères permettant de mesurer la dissimilarité entre un résultat de segmentation en régions/classes par rapport à un résultat de segmentation de référence.

Yasnoff et al. [Yasnoff *et al.* 1977] ont proposé une série de mesures d'erreur de classification lorsque les classes sont supposées connues, c'est à dire qu'il faut connaître pour chaque classe d'un résultat de segmentation la classe qui doit lui être associée dans un autre résultat de segmentation. Ces indicateurs permettent de rendre compte classe par classe des erreurs de classification, mais ne tiennent pas compte des informations spatiales sur les pixels mal classés. Un autre critère permettant de mesurer la dissimilarité entre deux résultats de segmentation est celui de Vinet [Vinet 1991] [Cocquerez et Philipp 1995]. Il consiste à apparier la région du résultat de segmentation avec une région de la vérité terrain maximisant le recouvrement des régions. Ce critère est facile à mettre en œuvre. Cependant, il n'apparie pas toutes les classes, donc il ne prend pas en compte toute l'information (entre autre la dispersion spatiale des pixels). Il part de l'hypothèse que deux classes sont à apparier si elles ont un ensemble maximal commun de pixels. Cette hypothèse est restrictive et privilégie les grandes régions. D'autres mesures d'erreur, basées sur le principe de recouvrement entre deux résultats de segmentation en régions ou classes, ont été définies par Hamming [Huang et Dom 1995], Kara-Falah [Kara-Falah 1995] et Martin [Martin *et al.* 2001].

4.3.1.3 Discussion

L'utilisation d'une connaissance rend cette approche potentiellement plus générale et moins biaisée. Cependant, son inconvénient majeur est le besoin d'une référence qui n'est pas toujours disponible. Nous voyons qu'il existe de nombreux critères d'évaluation supervisée dans la littérature. Il est néanmoins difficile de choisir parmi tous ces critères, ceux qu'il faut utiliser pour l'évaluation d'un résultat à partir d'une vérité terrain synthétique ou extraite par un expert. La raison essentielle vient de la méconnaissance de leurs performances.

4.3.2 Evaluation non supervisée

Les critères d'évaluation non supervisée permettent d'évaluer la qualité de résultats de segmentation d'une image sans aucune information (voir figure FIG. 4.6). Ils s'adressent donc plus particulièrement aux deux catégories de méthodes de segmentation nécessitant peu de connaissances *a priori* afin d'éviter l'introduction d'un biais dans l'évaluation. Il faut noter que ces critères mesurent en quelque sorte la cohérence du résultat de segmentation en se basant sur des conditions nécessaires que doit respecter toute méthode de segmentation.

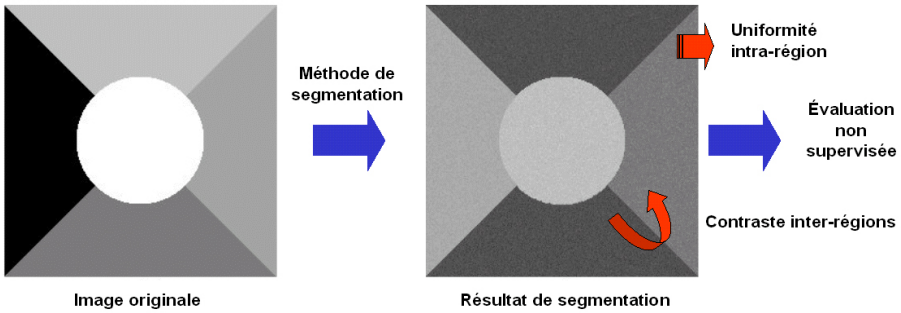


FIG. 4.6 – Évaluation non supervisée d'un résultat de segmentation à l'aide de calculs statistiques.

4.3.2.1 Évaluation de la détection de frontières

Kitchen et Rosenfeld [Kitchen et Rosenfeld 1984] ont proposé un critère d'évaluation de cartes de frontières fondée sur la *cohérence locale des frontières*. Cette cohérence est évaluée sur la base de deux critères de caractérisation des frontières : la continuité et la minceur. Les frontières sont supposées être générées par un détecteur de frontières de type gradient, suivi par un seuillage. Des critères définis par Levine et Nazif [Levine et Nazif 85] permettent d'évaluer deux types de frontières : les frontières séparant deux objets de la scène et les lignes qui traversent un objet, sans qu'il y ait de différence de couleur ou de texture entre les régions autour de cette ligne. Le premier type de frontière est évalué par le contraste entre les régions, alors que le second est évalué par le gradient moyen le long de la ligne. Dans le cas d'une image couleur, les calculs sont effectués séparément sur chaque canal et fournissent un vecteur de mesures. Han et Kim [Han et Kim 2002] ont défini une distance d'ambiguïté pouvant être utilisée dans la détermination de la justesse de détection des frontières. Trois mesures sont utilisées pour calculer l'ambiguïté : l'existence, la localisation et la disposition. L'ambiguïté d'existence et l'ambiguïté de localisation sont dérivées d'une généralisation de la représentation des frontières en utilisant des ensembles flous.

4.3.2.2 Évaluation de la segmentation en régions

L'un des premiers critères pouvant servir à qualifier un résultat de segmentation en régions est l'uniformité intra-région. En effet, sur une image faiblement ou non texturée, les régions segmentées doivent apparaître comme les plus homogènes possibles. Ainsi, Weszka et Rosenfeld [Weszka et Rosenfeld 1978] ont proposé un critère d'évaluation avec seuillage qui utilise une mesure de bruit. L'image de départ doit comporter des objets de forme compacte et un fond faiblement texturé. La quantité de bruit pour une image seuillée est calculée en utilisant la matrice de co-occurrence de niveaux de gris de l'image. Les éléments de la matrice de co-occurrence représentant le pourcentage d'adjacence entre les objets et le fond sont additionnés. Le résultat indique alors une quantité de bruit. Moins il y a de bruit, plus le résultat de segmentation est supposé être

bon. Avec la même idée d'uniformité intra-région, Levine et Nazif [Levine et Nazif 85] ont défini un critère. Cependant, contrairement à Weszka et Rosenfeld qui prenaient en compte l'image entière, l'idée est ici de calculer l'uniformité d'une caractéristique sur une région en se basant sur la variance de cette caractéristique. Reposant sur le même principe, la mesure d'homogénéité de Cochran [Cocquerez et Devars 1985] donne un indice de confiance sur l'homogénéité des régions.

D'un point de vue complémentaire à l'uniformité intra-région, un autre critère venant rapidement à l'esprit pour évaluer un résultat de segmentation est le contraste inter-régions. En effet, deux régions adjacentes sont supposées avoir un contenu différent. Levine et Nazif [Levine et Nazif 85] ont défini la mesure de contraste en niveaux de gris entre deux régions. Les deux catégories de critères Intra et Inter-régions que nous venons de voir permettent de mesurer deux aspects différents d'un résultat de segmentation d'une image. Quelques chercheurs ont eu l'idée par la suite de combiner ces deux catégories de critères. Ainsi, Zéboudj [Zeboudj 1988] propose une mesure reposant sur les principes conjugués de contraste inter-régions maximal et intra-région minimal mesurés sur le voisinage d'un pixel.

Les critères énoncés jusqu'à présent étaient tous orientés vers un aspect particulier du résultat de segmentation à évaluer. Les critères présentés dans ce paragraphe regroupent plusieurs de ces aspects de manière à en tirer le meilleur profit. Ainsi, Liu et Yang [Liu et Yang 1994] ont proposé une fonction élaborée pour incorporer trois des quatre critères heuristiques suggérés par Haralick et Shapiro [Haralick et Shapiro 1985] pour définir une bonne segmentation. D'après les auteurs, ces critères ne nécessitent pas la définition préalable d'une valeur de seuillage pour les propriétés subjectives telles que la taille, la forme ou l'homogénéité des régions. Borsotti et al. [Borsotti *et al.* 1998] ont identifié des limitations du critère précédent et l'ont alors modifié. L'objectif était de pénaliser les résultats de segmentation présentant beaucoup de petites régions ainsi que des régions non homogènes. Un autre critère d'évaluation non supervisée de résultats de segmentation reposant sur une approche multi-critères de qualité a été proposé par Philipp-Foliguet et Guigues dans [Philipp-Foliguet et Guigues 2005]. Le principe utilisé dans ce critère d'évaluation est de considérer tout d'abord le problème de segmentation d'images comme un problème de modélisation par morceaux d'une image : modélisation constante, polynômiale, gaussienne, lisse par morceaux... chaque région correspondant à un "morceau" du modèle. Alors, une énergie peut être calculée sur l'ensemble du modèle permettant ainsi d'évaluer la qualité du résultat de segmentation.

Le critère que j'ai proposé en 1999 dans ma thèse de doctorat [Rosenberger 1999] est basé sur un calcul d'homogénéité intra-classe et de contraste inter-classes. Ce critère a la particularité de prendre en compte le caractère uniforme ou texturé des régions détectées par segmentation. En effet, la plupart des critères de la littérature font l'hypothèse que l'image est uniforme ou faiblement texturée. Or, les images naturelles sont rarement uniformes. La fonction quantifiant la qualité d'un résultat de segmentation I_S est définie

comme suit :

$$ROS(I_S) = \frac{C_{Intra}(I_S) + 1 - C_{Inter}(I_S)}{2} \tag{4.1}$$

où
- $C_{Intra}(I_S)$ correspond à un calcul de disparité intra-classe globale quantifiant l'homogénéité des régions du résultat I_S,
- $C_{Inter}(I_S)$ correspond à la disparité inter-classes globale mesurant la disparité globale des régions du résultat I_S.

La disparité $C_{Intra}(I_S)$ d'un résultat de segmentation I_S d'une image I traduit la stabilité des statistiques des régions de l'image. Elle est calculée à partir de la disparité intra-classe $C_{Intra}(R_k)$ des différentes régions $R_k,\ k = 1..NR$, du résultat de segmentation I_S :

$$C_{Intra}(I_S) = \frac{1}{NR} \sum_{k=1}^{NR} \frac{Card(R_k)}{Card(I)} C_{Intra}(R_k) \tag{4.2}$$

L'importance de la disparité intra-classe d'une région R_k dans le calcul de la disparité intra-classe globale est proportionnelle au nombre de pixels $Card(R_k)$ de cette région. Ainsi, une région de petite taille possède une influence moindre dans le calcul de la disparité intra-classe globale. Enfin, une région contenant deux primitives différentes possède alors une disparité intra-classe élevée, une primitive correspondant ici à un ensemble de caractéristiques permettant de définir les similitudes entre régions [Rosenberger 1999].

Le principe est identique pour le calcul de la disparité inter-classes globale $C_{Inter}(I_S)$ du résultat I_S qui mesure la disparité de chaque région avec ses voisines :

$$C_{Inter}(I_S) = \frac{1}{NR} \sum_{k=1}^{NR} \frac{Card(R_k)}{Card(I)} C_{Inter}(R_k) \tag{4.3}$$

Dans les calculs de disparité intra et inter-classes, la nature de la région segmentée est prise en compte, à savoir uniforme ou texturée. Un critère de décision amélioré dans [Rosenberger et Chehdi 2003a] permet de déterminer l'appartenance d'une région à l'une de ces deux catégories.

Dans le cas d'une région uniforme, la disparité intra-classe est donnée par l'écart-type normalisé de celle ci. Cette caractéristique d'ordre 2 sur la dispersion des niveaux de gris d'une région est suffisante pour caractériser la disparité intra-classe d'une classe uniforme. Si la région est texturée, l'écart type ne donne pas d'information suffisante sur l'homogénéité de celle-ci. La région considérée est analysée en calculant sur une fenêtre glissante (voir figure FIG. 4.7) des attributs de texture (cooccurrence, moments...). Une mesure de dispersion de ces attributs de texture permet de quantifier l'homogénéité de la

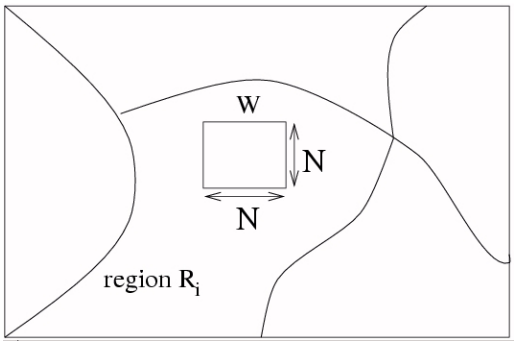


FIG. 4.7 – Découpage en blocs du résultat de segmentation à évaluer.

région.

Le procédé d'évaluation de la disparité inter-classes d'une région est similaire à ce qui a été réalisé précédemment. La disparité inter-classes s'écrit dans ce cas, comme la moyenne de la disparité d'une région avec ses voisines.

Pour le calcul de la disparité entre régions de types différents, elle est mise à 1 et est donc rendue maximale. Une région est décrite par la valeur moyenne des statistiques représentatives de la primitive. Ainsi, une région uniforme est décrite par sa moyenne de niveaux de gris et une région texturée par le vecteur d'attributs de texture moyen, c'est à dire son barycentre. La disparité inter-classes est alors calculée en considérant l'écart entre ces valeurs moyennes. L'originalité de ce critère réside dans sa capacité à évaluer des résultats de segmentation d'images texturées.

4.3.2.3 Discussion

Les critères d'évaluation non supervisée possèdent un avantage majeur : ils ne requièrent qu'une définition par l'utilisateur d'une métrique de mesure de qualité. Ainsi, ils sont totalement automatisables et ne nécessitent pas l'intervention d'un expert. Il ne faut pas oublier que ce type de critère ne peut s'adresser qu'à des méthodes de segmentation *"non supervisée"* ou *"semi-supervisée"* (d'après la classification de la section 4.2).

Cependant, du fait de l'autonomie de ces critères d'évaluation, il est d'autant plus difficile d'évaluer correctement leur comportement face à des situations diverses. Le défaut principal de ces critères d'évaluation réside dans le fait que la majorité d'entre eux donne de mauvais résultats sur des images texturées. Un autre point négatif de ces critères est que les mesures de qualité métriques sont au mieux heuristiques, car aucune connaissance sur les éléments à segmenter n'est disponible. De plus, ces critères peuvent présenter un fort biais face à certaines méthodes de segmentation qui vont utiliser la même technique que celle prise en compte par l'évaluation. Par exemple, un critère d'évaluation reposant sur le critère d'uniformité intra-région va avoir un fort biais lors de l'évaluation d'une

méthode de segmentation telle qu'une croissance de régions.

4.4 Notre contribution

Notre contribution à ce sujet se situe à différents niveaux :
– étude de l'efficacité des critères existants dans la littérature,
– proposition d'un ensemble de critères d'évaluation dans le cas non supervisé,
– exploitation des critères d'évaluation pour différentes applications.

4.4.1 Comparaison de critères d'évaluation

Nous avons étudié l'efficacité des critères d'évaluation de l'état de l'art de façon empirique et en respectant un protocole expérimental rigoureux.

4.4.1.1 Evaluation supervisée

Nous avons étudié les techniques d'évaluation supervisée suivant trois approches :
– similarité de frontières,
– jugement expert,
– similarité de la partition en régions.

4.4.1.1.1 Approche frontière

Les critères d'évaluation supervisée de résultats de segmentation en frontières sont généralement basés sur le calcul de distances entre un résultat de segmentation et une référence connue, que l'on nomme vérité terrain. Afin d'évaluer des résultats de segmentation, nous avons sélectionné 14 critères de l'état de l'art : sur-détection SUR, sous-détection $SOUS$, localisation LOC, distances Lq ($Lq1$ à $Lq4$), distance de Küllback DKu, distance de Bhattacharyya DBh, distance de Jensen DJe, indice de Pratt PRA, distance de Hausdorff HAU, critère d'Odet de sur-segmentation ODI et critère d'Odet de sous-segmentation UDI.

Dans un premier temps, nous avons comparé l'évolution de ces critères d'évaluation sur des résultats de segmentation synthétiques. Cela nous a permis de mesurer leur pertinence vis à vis d'un biais que nous maîtrisons. Les études menées portent successivement sur l'influence d'une erreur de localisation, de la sous-segmentation, de la sur-segmentation, puis la combinaison de ces deux dernières altérations. La figure FIG. 4.8 montre des exemples de dégradations d'une vérité terrain.

Nous avons analysé le comportement de tous les critères face à ces dégradations. A titre d'exemple, la figure FIG. 4.9 présente l'évolution des critères PRA, HAU et ODI face à l'ajout de bruit impulsionnel à la vérité terrain. Nous simulons dans ce cas, une sur-segmentation de l'image pouvant se produire par exemple dans le cas

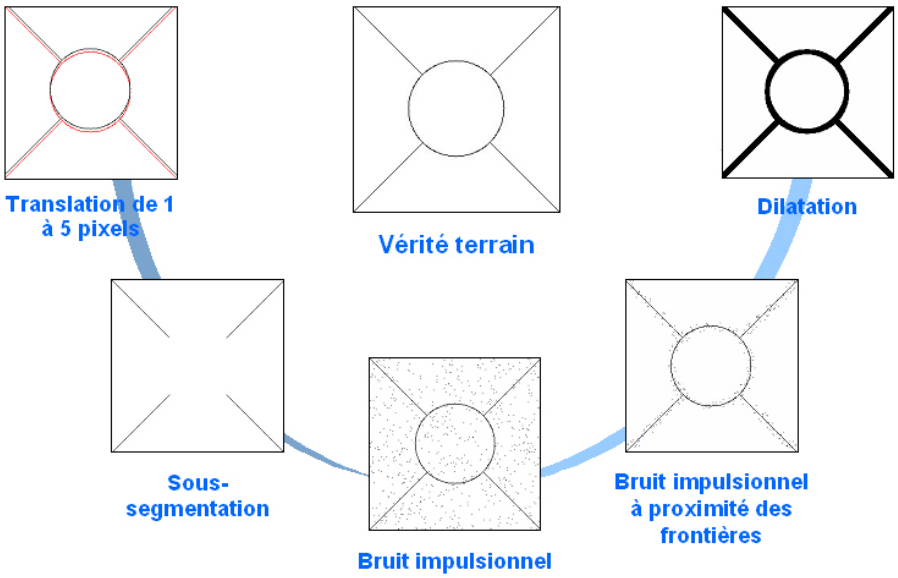


FIG. 4.8 – Plusieurs dégradations synthétiques d'une vérité terrain en contour.

d'une image texturée. Afin de faciliter l'analyse des résultats, nous avons normalisé la valeur des critères d'évaluation de telle façon qu'elle soit minimale pour un résultat de segmentation très proche de la vérité terrain. Le résultat attendu pour cette expérience est que la valeur d'un critère d'évaluation augmente avec l'ajout de bruit (simulant ici une sur-segmentation). Ainsi, dans la figure FIG. 4.9, seul le critère PRA pénalise de façon correcte l'ajout de bruit sur la vérité terrain.

Cette étude sur des résultats synthétiques montre que la majorité des critères étudiés se comporte correctement. Les critères ayant eu un comportement général correct sur nos bases de résultats de segmentation synthétiques sont : LOC, DBh, DJe, DKu, PRA, $Lq1$, $Lq2$, $Lq3$, $Lq4$. L'ensemble de ces critères évolue en général de la même manière à un facteur d'échelle près, exception faite de PRA qui donne des avis plus tranchés.

Une étude comparative sur des résultats de segmentation réels d'images naturelles a été réalisée [Chabrier *et al.* 2004b]. La base utilisée est composée de 300 images naturelles issues de la base ©Corel. Des résultats de segmentation réalisés manuellement par des étudiants ont été utilisés[1]. La figure FIG. 4.10 représente quelques images de cette base ainsi que des résultats de segmentation réalisés manuellement.

Compte tenu de la diversité de ces résultats de segmentation manuelle, nous les avons fusionnés afin d'obtenir pour chaque image de la base un résultat compromis, considéré comme une vérité terrain [Chabrier *et al.* 2004b]. Chaque image de la base a été segmentée par 10 méthodes de détection de contours dont le gradient de luminosité

[1]http ://www.cs.berkeley.edu/projects/vision/grouping/segbench/

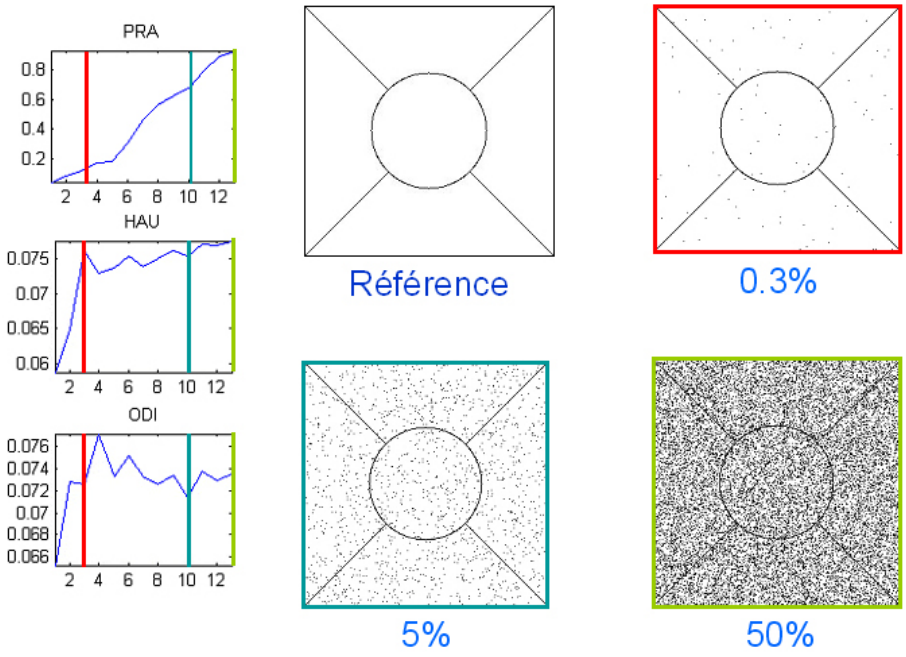


FIG. 4.9 – Evolution de trois critères d'évaluation de détection de contours face à l'ajout de bruit impulsionnel sur la vérité terrain (l'abscisse des courbes correspondent aux différents résultats synthétiques de segmentation et l'ordonnée à la valeur de chaque critère).

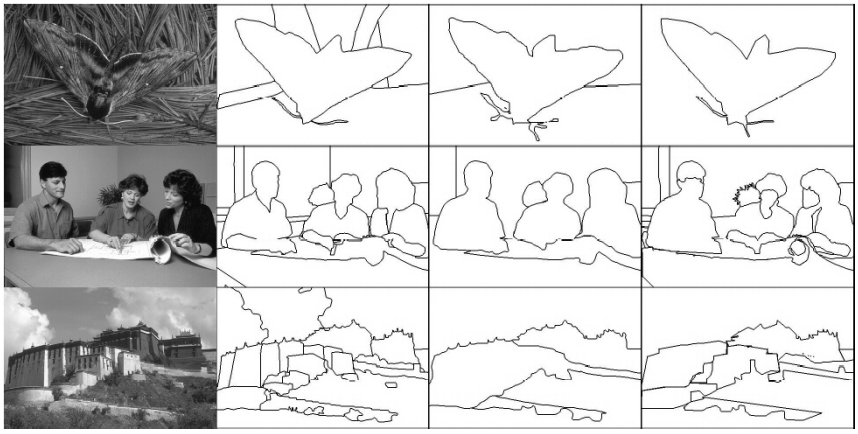


FIG. 4.10 – Différents résultats de segmentation extraits manuellement.

ou le filtre de Canny. Trois résultats de segmentation ont été obtenus pour différents seuils de détection. Un critère que nous avons défini de façon empirique [Chabrier 2005] permet d'obtenir un résultat de segmentation compromis à partir de la vérité terrain. Deux autres résultats de segmentation correspondent à une sous et à une sur segmentation de ce résultat. La figure FIG. 4.11 présente les résultats de segmentation obtenus sur une image de la base Corel.

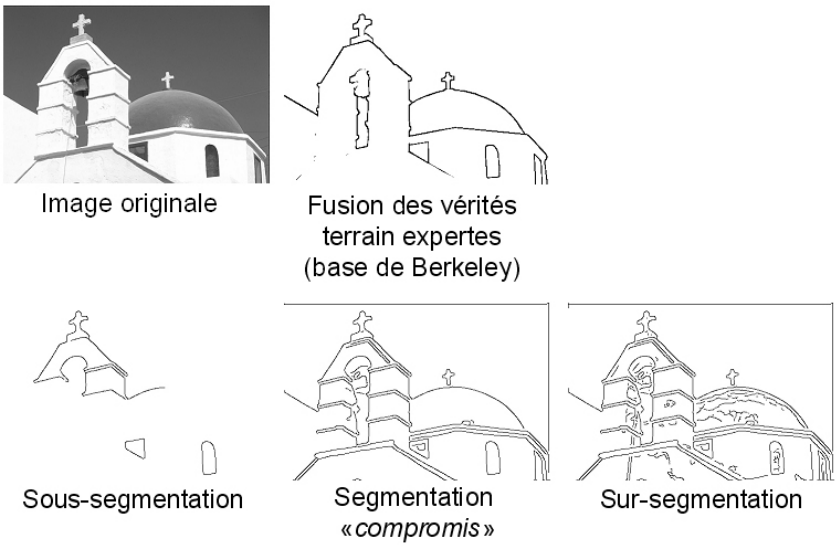


FIG. 4.11 – Résultats de segmentation utilisés pour la comparaison de critères d'évaluation supervisée.

Telle que nous avons conçu notre expérience, un critère d'évaluation pertinent doit préférer le résultat compromis aux deux autres. Chaque résultat de segmentation est classé par tous les critères tel que le classement 1 corresponde au meilleur résultat de segmentation d'après le critère considéré. La moyenne de classement obtenue sur l'ensemble des résultats est présentée pour chaque critère dans le tableau TAB. 4.1.

Tous les critères, sauf un, ont voté pour un résultat sous ou sur-segmenté comme étant le meilleur. La conception même de certains critères le justifie. Les critères LOC, Dbh, Dje, Dku, HAU, $Lq1$, $Lq2$, $Lq3$ et $Lq4$ calculent bien une distance entre deux cartes de frontières, cependant, ils pénalisent grandement les résultats sur-segmentés comme nous pouvons le constater dans le tableau TAB. 4.1 où ils choisissent dans presque tous les cas les résultats sous-segmentés. Seul le critère PRA a su trouver dans la majorité des cas (89%) le résultat de segmentation dit de *Compromis* comme étant le meilleur. Ce critère est donc le plus performant parmi tous ceux que nous avons testés.

Le critère que nous avons utilisé pour déterminer le résultat compromis a incontestablement un impact sur le résultat de la comparaison. Une étude est actuellement en cours afin d'identifier le biais engendré par le critère compromis. Le résultat de segmentation compromis doit être néanmoins plus proche de la valeur optimale d'un critère d'évaluation pertinent que les deux autres résultats. Il est quoi qu'il arrive nécessaire d'en utiliser un pour définir la référence. Les résultats obtenus dans le cas d'images naturelles confortent ceux obtenus dans le cas synthétique.

Critère	Résultat Sous-segmenté	Résultat Compromis	Résultat Sur-segmenté
LOC	1.1	2.0	3.0
SOUS	2.8	1.9	1.0
SUR	1.1	2.0	3.0
DBh	1.1	2.0	3.0
DJe	1.1	2.0	3.0
DKu	1.1	2.0	3.0
PRA	1.7	**1.3**	2.8
HAU	1.4	2.0	2.9
Lq1	1.1	2.0	3.0
Lq2	1.1	2.0	3.0
Lq3	1.1	2.0	3.0
Lq4	1.1	2.0	3.0
ODI	1.3	1.9	2.7
UDI	2.8	1.9	1.0

TAB. 4.1 – Classement moyen de chaque résultat de segmentation par les différents critères, 1 indiquant le meilleur résultat.

4.4.1.1.2 Jugement expert

Nous avons réalisé une étude impliquant le jugement d'experts pour la comparaison de résultats de segmentation. Cette étude a été réalisée grâce au concours de la communauté française du traitement d'images via la liste de diffusion du groupe GDR ISIS. Nous avons demandé aux participants de classer 5 résultats de segmentation d'une image. Les images sont assez simples et l'utilisation d'une méthode de segmentation semi-supervisée (en définissant le niveau de précision adapté en fonction de l'image) permet d'obtenir un résultat proche de l'interprétation. Les résultats de segmentation ont été obtenus par différentes méthodes de segmentation comme par exemple la méthode EDISON. Une méthode d'appariement de couleur a été utilisée afin de faciliter visuellement la comparaison des résultats de segmentation [Chabrier *et al.* 2004b]. La figure FIG. 4.12 présente l'interface utilisée par les utilisateurs.

Au final, 160 personnes ont participé à ce test. L'écart type des résultats donnés par les experts est de l'ordre de 0.65 mettant en évidence la cohérence des résultats. Contrairement à ce qui a été présenté dans le chapitre précédent, aucun contrôle sur l'environnement de l'expert n'a été mis en place. Néanmoins, dans la mesure où les résultats de segmentation étaient présentés en couleur, cela se justifiait moins.

Cette approche a donné des résultats intéressants même si les images utilisées étaient assez simples donc peu représentatives de la complexité des images naturelles. Cette étude a été exploitée afin d'obtenir une information sur la qualité de résultats de segmentation d'images en région.

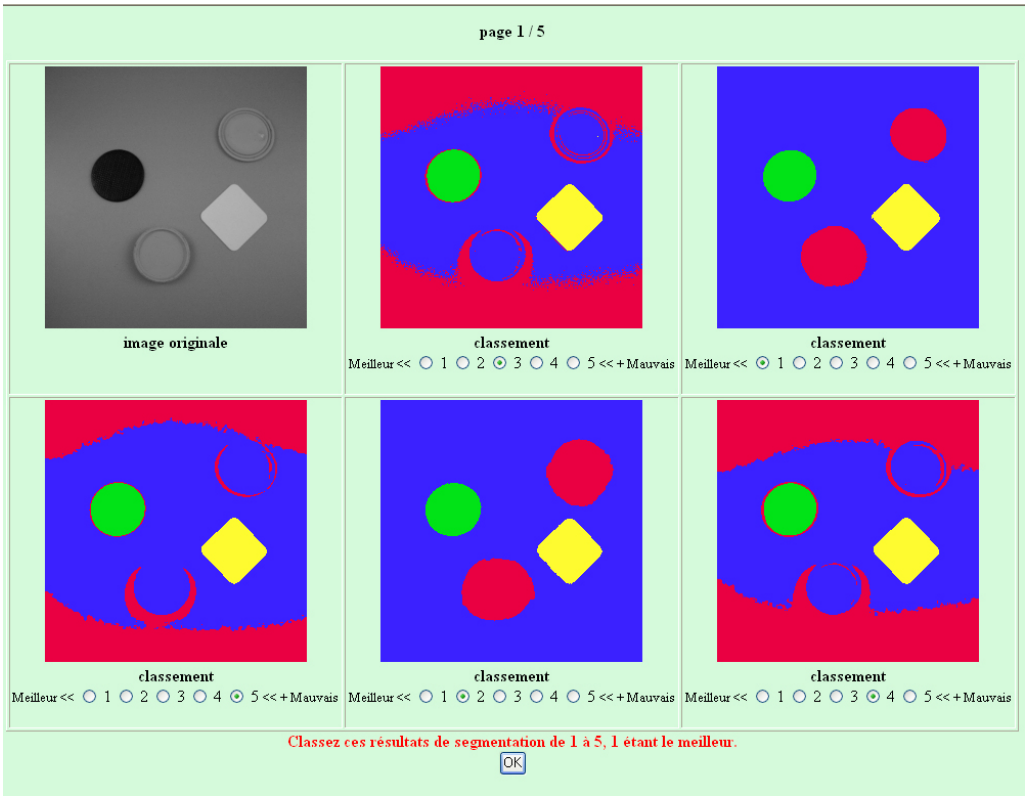


FIG. 4.12 – Interface du test pour la comparaison de résultats de segmentation.

4.4.1.1.3 Approche région

Nous avons procédé de manière équivalente à l'approche frontière. Les critères d'évaluation retenus pour cette étude ont été présentés dans l'état de l'art : VIN de Vinet [Vinet 1991], $HAM1$ de Hamming [Huang et Dom 1995], $YAS1$, $YAS2$, $YAS3$ de Yasnoff [Yasnoff *et al.* 1977], $MAR1$ et $MAR2$ de Martin [Martin *et al.* 2001].

Une étude sur des résultats de segmentation synthétiques a tout d'abord été réalisée. Les études menées portent successivement sur l'influence de l'erreur de localisation, de la sous-segmentation, de la sur-segmentation, puis la combinaison de ces deux dernières altérations. La figure FIG. 4.13 montre des exemples de dégradations d'une vérité terrain.

Nous avons analysé le comportement de tous les critères face à ces dégradations. A titre d'exemple, la figure FIG. 4.14 présente l'évolution des critères VIN, $YAS1$ et $YAS3$ face à la sous-segmentation. De la même manière que dans le cas frontière, nous avons normalisé la valeur des critères d'évaluation de telle façon qu'elle soit minimale pour un résultat de segmentation très proche de la vérité terrain. La figure FIG. 4.14 illustre l'évolution correcte des critères VIN et $YAS1$ face à des résultats sous-segmentés comparée à celle de $YAS3$. En effet, ces deux critères pénalise correctement la sous-segmentation (en ayant une valeur plus grande) avec un comportement plus fin du

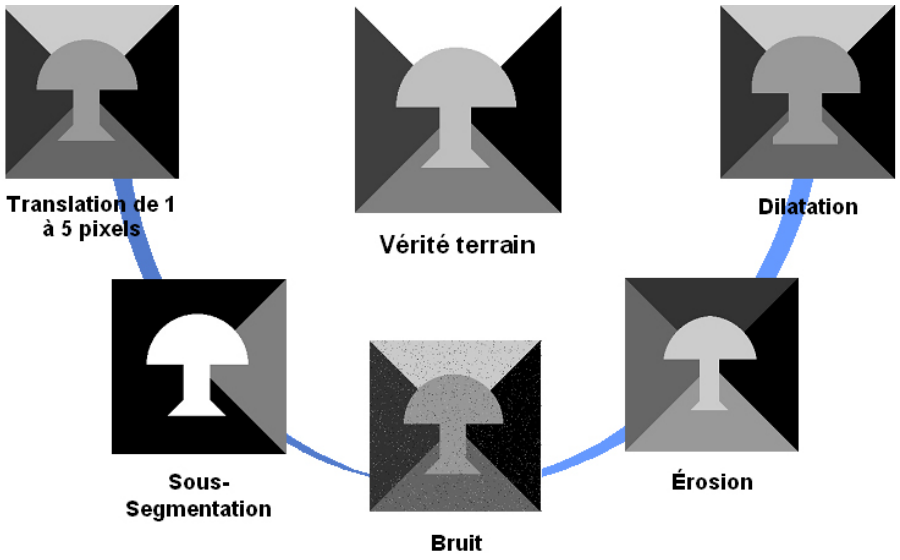


FIG. 4.13 – Plusieurs dégradations synthétiques d'une vérité terrain en régions.

critère VIN distinguant des sous-segmentations de l'image de même nombre de classes (qui sont de taille différentes).

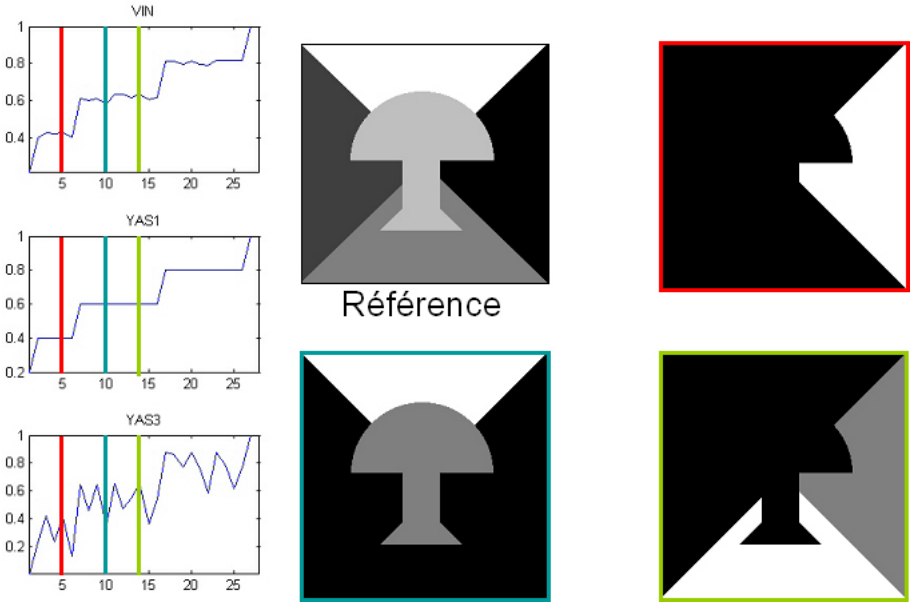


FIG. 4.14 – Evolution de trois critères d'évaluation région face à la sous-segmentation (l'abscisse des courbes correspondent aux différents résultats synthétiques de segmentation et l'ordonnée à la valeur de chaque critère).

A l'issue de toutes les expérimentations, les critères ayant eu un comportement

général correct sur les bases de résultats de segmentation synthétiques sont : VIN, $HAM1$, $YAS2$, $MAR1$ et $MAR2$. Dans l'ensemble, les critères les plus discriminants entre deux résultats de segmentation proches sont les critères VIN et $MAR1$.

Le deuxième type d'expérimentation est la comparaison de ces critères sur des résultats de segmentation d'images naturelles. Pour ce faire, nous avons exploité les résultats de l'étude utilisant le jugement expert. Nous avions des images naturelles avec 5 résultats de segmentation classés par des experts. Nous avons considéré le meilleur résultat classé par les experts (score moyen le plus proche de 1) comme une vérité terrain. Les 4 résultats de segmentation restants ont été comparés avec la vérité terrain à l'aide des différents critères d'évaluation. Afin d'estimer la pertinence de jugement du critère d'évaluation, nous avons calculé le taux de bonne comparaison (noté TBC et défini dans le chapitre précédent) entre le jugement expert et celui obtenu par l'utilisation d'un critère. A partir de ces données, nous obtenons les résultats présentés dans le tableau TAB. 4.2 indiquant le TBC de chaque critère d'évaluation. Nous pouvons constater que les taux de bonne comparaison sont au mieux de l'ordre de 86% en ce qui concerne le critère VIN. Tous les autres critères ont un taux de bonne comparaison de l'ordre de 72%. Ainsi, le critère VIN se révèle être le critère le plus intéressant.

	VIN	HAM	YAS1	YAS2	YAS3	MAR1	MAR2
Taux de Similarité	86%	70%	72%	72%	74%	72%	72%

TAB. 4.2 – Présentation du taux de bonne comparaison des critères d'évaluation supervisée de résultats de segmentation en régions par rapport au classement d'experts.

4.4.1.2 Evaluation non supervisée

Nous avons étudié des critères d'évaluation non supervisée de l'état de l'art. Dans ce cas, aucune connaissance *a priori* n'est utilisée ([Laurent *et al.* 2003, Chabrier *et al.* 2006a]). Par conséquent, l'utilisation de ces critères n'est pas pertinente dans le cas d'une segmentation supervisée dans la mesure où il n'utilise aucune information *a priori*. Nous n'avons abordé que le cas des résultats de segmentation en régions (et plus particulièrement en classes). En effet, les critères d'évaluation de résultats de segmentation dans le cas de la détection de frontières d'un point de vue non supervisée sont peu nombreux et de surcroît très mal adaptés dans le cas des images texturées.

Nous avons créé une base d'images synthétiques dénoté BCU comprenant 14 sous-bases d'images possédant respectivement de 2 à 15 classes. La figure FIG. 4.15 présente les vérités terrain qui ont permis de créer chaque image des 14 sous-bases. Chaque sous-base a donc un nombre de classes fixé et est composée de 600 images créées de la manière suivante :

– $BC100U$: 100 images sont composées de 100% de classes uniformes,

- $BC75U$: 100 images sont composées de 75% de classes uniformes et de 25% de classes texturées,
- $BC50U$: 100 images sont composées de 50% de classes uniformes et de 50% de classes texturées,
- $BC25U$: 100 images sont composées de 25% de classes uniformes et de 75% de classes texturées,
- $BC0U$: 100 images sont composées de 100% de classes texturées,
- $BC0UN$: 100 images sont composées de 100% de classes texturées de niveaux de gris moyens identiques pour une même image (les moyennes de niveaux de gris vont de 75 à 175 pour les 100 images de la base). Ce dernier ensemble d'images texturées est présent afin de permettre l'étude de l'impact de classes texturées non différenciables par une statistique d'ordre 1.

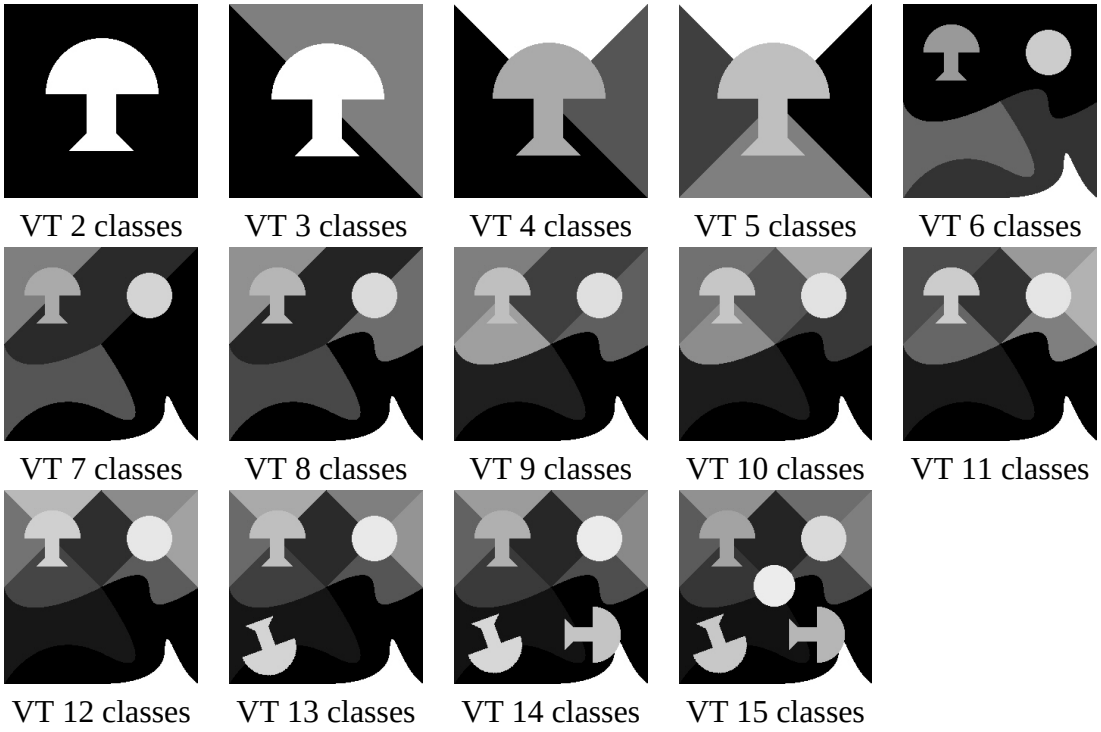


FIG. 4.15 – Vérités-terrain utilisées pour la création des bases d'images synthétiques de 2 à 15 classes.

Nous avons utilisé 3 méthodes de segmentation (en fait des méthodes de classification) : un partitionnement flou (FCM) [Bezdeck 1981], une relaxation de ce résultat et la méthode Edison [Comaniciu et Meer 2002]. La vérité terrain constitue un quatrième résultat considéré comme idéal. La figure FIG. 4.16 présente un exemple de résultats de segmentation par ces méthodes sur une image de la base.

Les critères utilisés pour évaluer ces bases de résultats de segmentation ont été

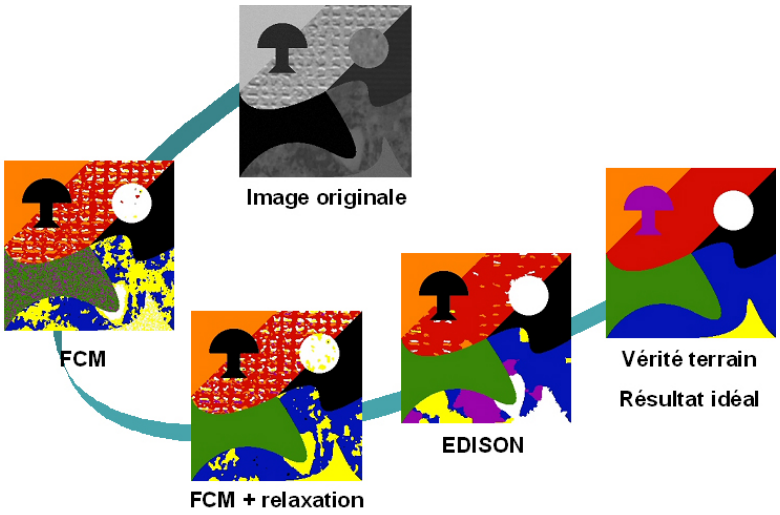


FIG. 4.16 – Exemple d'une image comportant 5 classes et ses résultats de segmentation avec couleurs appariées.

choisis parmi ceux présentés dans l'état de l'art : $LEV1$ [Levine et Nazif 85], $LEV3$ [Levine et Nazif 85], $LEV4$ [Levine et Nazif 85], BOR [Borsotti *et al.* 1998], ZEB [Zeboudj 1988] et ROS [Rosenberger 1999]. Afin de simplifier la comparaison des critères, ceux-ci ont été normalisés pour évoluer entre 0 et 1, 1 correspondant à une qualité optimale.

Nous avons calculé les critères d'évaluation présentés ci-dessus sur la base de résultats de segmentation BCU. Nous disposons alors de 8400 images originales, de 8400×3 résultats de segmentation calculés et de 14 vérités terrain servant de 4ème résultat de segmentation, soit un total de 33600 résultats de segmentation différents. Comme nous l'avons vu dans la section précédente, le critère VIN [Vinet 1991] s'est avéré être le plus efficace en supervisé, nous avons donc décidé de prendre ce critère comme référence. Nous avons calculé ce critère de référence sur l'ensemble des résultats de segmentation des différentes bases par rapport à la vérité terrain associée. Notre objectif est alors de comparer le classement des différents résultats de segmentation pour chaque image par les critères d'évaluation non supervisée par rapport au classement établi par VIN.

Nous utilisons alors le TBC comme indicateur de performance, que nous avons calculé sur l'ensemble des résultats d'évaluation des 6 différents critères pour chaque base. Afin de simplifier les notations, nous utiliserons par la suite la notation TBC_{VIN} pour représenter le calcul du TBC par rapport aux résultats du critère VIN sur une base de résultats de segmentation particulière. Le tableau TAB. 4.3 présente une synthèse des résultats sur l'ensemble des bases et nous donne donc un aperçu global du comportement des différents critères en fonction du type des images traitées.

	ROS	ZEB	BOR	$LEV3$	$LEV1$	$LEV4$
BCU	57.7%	49.4%	24.5%	**66.1%**	31.3%	36.6%
$BC100U$	50.7%	**88.4%**	65.7%	73.7%	52.2%	66%
$BC75U$	52.9%	67.3%	27.5%	**69.9%**	40.8%	40%
$BC50U$	55.8%	54.5%	19.2%	**71.8%**	33.5%	32.2%
$BC25U$	60.8%	38.8%	12.5%	**72.8%**	25.7%	25.8%
$BC0U$	65%	32.2%	11.1%	**74.6%**	20%	23.5%
$BC0UN$	**61.3%**	15.1%	11.2%	33.6%	15.7%	32.3%

TAB. 4.3 – TBC des critères d'évaluation non supervisée par rapport au critère VIN sur différentes sous-bases d'images.

Dans le cas d'images totalement uniformes, le critère ZEB s'est avéré être le plus intéressant pour évaluer les résultats de segmentation avec un taux de bonne comparaison par rapport à notre critère de référence (VIN) de 88% en moyenne. La texture occupe en général une grande partie des images naturelles, et c'est pourquoi, nous avons étudié l'impact de l'augmentation de la proportion de classes texturées dans les images synthétiques. Il en résulte que la fiabilité du critère de Zéboudj décroît progressivement au fur et à mesure de ces ajouts. Le critère $LEV3$ au contraire possède un TBC moyen de l'ordre de 70%, constant au fur et à mesure de l'augmentation de la proportion de classes texturées à l'exception de l'étude sur la base $BC0UN$. Ces résultats peuvent paraître étonnants dans la mesure où ce critère ne tient absolument pas compte de la texture. Dans le cas où toutes les classes possèdent la même moyenne de niveau de gris, il n'obtient qu'un TBC_{VIN} moyen de l'ordre de 30%. Ceci peut s'expliquer par le fait qu'il est très rare que deux régions limitrophes dans un résultat de segmentation possèdent une moyenne de niveau de gris proche. Le critère ROS, quant à lui, possède un taux de classification correcte qui augmente au fur et à mesure des ajouts de textures, pour atteindre 60% en moyenne dans le cas d'évaluation de résultats de segmentation d'images fortement texturées. Ses résultats restent stables même si la moyenne de niveaux de gris des classes voisines est identique, au contraire du critère $LEV3$.

4.4.1.3 Discussion

Dans le cas de l'évaluation supervisée, deux critères se sont révélés être, à l'issue de nos expérimentations, les plus performants : le critère de Pratt PRA pour les résultats de détection de contours et le critère de Vinet VIB pour l'approche région.

En ce qui concerne l'évaluation non supervisée, trois critères ont montré leur intérêt en fonction du type de l'image originale : le critère de Zeboudj ZEB pour des images uniformes, le critère ROS que j'ai proposé durant ma thèse dans le cas texturé et le critère de Levine et Nazif $LEV3$ dans le cas général.

On peut tirer divers enseignements de ces expérimentations. D'une part, les critères d'évaluation supervisée permettent un jugement plus pertinent si l'on considère le critère de comparaison TBC. Ceci était attendu puisque ces critères exploitent une information supplémentaire comme une vérité terrain. Par ailleurs, le type de l'image originale influe grandement sur la performance des critères d'évaluation non supervisée. Il ne faut pas oublier que l'image originale est la seule information disponible pour ce type de critères. Enfin, l'analyse des performances de critères d'évaluation nécessite un grand nombre de données (images, résultats de segmentation, valeurs des critères d'évaluation) nécessitant plusieurs centaines d'heures de calcul.

4.4.2 Fusion de critères d'évaluation

L'objectif est ici de combiner par différentes méthodes l'information issue de plusieurs critères d'évaluation non supervisée (6 dans notre cas) pour la comparaison de 2 résultats de segmentation. Afin de paramétrer cette combinaison, nous avons utilisé la base de résultats de segmentation des images synthétiques utilisées lors de l'étude comparative. Nous avions donc 33600 résultats de segmentation d'images de qualité connue. Ceci fait un total de 50400 comparaisons (6 comparaisons à partir des 4 résultats de segmentation d'une image originale). Trois approches ont été utilisées pour réaliser cette fusion.

4.4.2.1 Critère adaptatif

Dans l'étude comparative des critères d'évaluation non supervisée, nous avons identifié les critères à utiliser en fonction du type d'image originale (uniforme, texturé ou bien mixte) dont on souhaite évaluer des résultats de segmentation. L'idée de ce premier critère est d'identifier le type de l'image originale afin de choisir le critère d'évaluation le plus pertinent [Chabrier *et al.* 2006b]. La figure FIG. 4.17 présente le principe de ce critère.

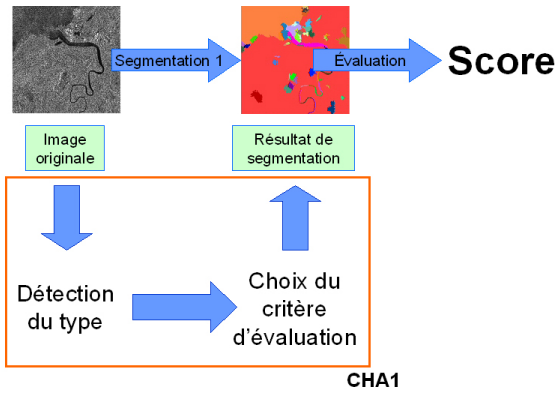


FIG. 4.17 – Principe du critère d'évaluation adaptatif.

L'identification du type de l'image originale est obtenue en calculant des attributs de texture (issus de la matrice de cooccurrence et les moments de niveaux de gris) sur l'image originale. Un apprentissage préalable est réalisé à partir de descripteurs calculés sur des images synthétiques. Nous avons utilisé un SVM pour les phases d'apprentissage et de reconnaissance. La méthode testée sur 8400 images synthétiques a permis une bonne identification du type de l'image dans 95% des cas.

En fonction du type d'image détecté, nous avons employé le critère d'évaluation correspondant. Le TBC_{VIN} global obtenu par ce nouveau critère adaptatif que nous appelons $CHA1$ est de 73.4% mettant en évidence un réel apport de la méthodologie puisque le critère le plus performant lors de notre étude comparative est $LEV3$ et permettait un taux de bonne comparaison avec la référence de 66%.

4.4.2.2 Fusion par combinaison linéaire

Nous avons adopté l'approche de fusion par combinaison linéaire des 6 critères d'évaluation (celle-ci est détaillée dans le chapitre précédent [Chabrier *et al.* 2005c]). Dans un premier temps, nous nous intéressons au gain obtenu par la fusion des 6 critères en utilisant une partie des résultats de notre base globale en apprentissage et le reste comme base de test. Ainsi, nous présentons dans la figure FIG. 4.18, le TBC_{VIN} obtenu pour différentes proportions de la base globale utilisée en apprentissage, avec les éléments restants utilisés pour tester le résultat de la fusion.

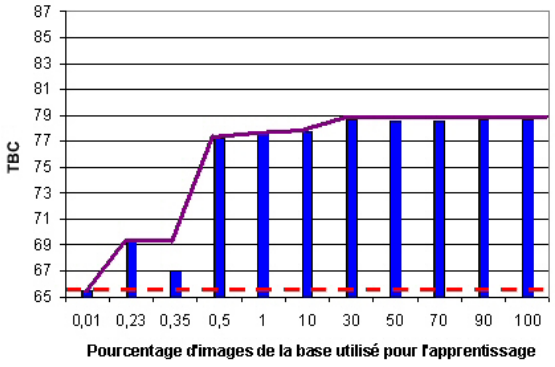


FIG. 4.18 – TBC_{VIN} de la fusion des 6 critères par l'algorithme génétique sur un pourcentage de la base globale (barres en bleu) par rapport à l'optimal obtenu sans fusion (pointillés rouges).

Lorsque nous prenons une image de la base comme apprentissage, cela signifie que nous utilisons les 6 résultats de comparaison obtenus à partir de 4 résultats de segmentation de cette image. La figure FIG. 4.18 présente ainsi les résultats de la fusion pour 1, 10, 20, 30, 42, ..., 8400 images de la base, ce qui correspond à 0.01% jusqu'à 100% de la base. Un apprentissage sur 30% des résultats de segmentation (10080 pour

être précis) permet d'obtenir une valeur du TBC proche de 79% mettant en évidence un gain notable par rapport au meilleur critère.

Nous avons ensuite réalisé la fusion par combinaison linéaire de n critères parmi les 6 choisis par ordre de performance de TBC_{VIN} dans le tableau TAB. 4.3, n allant de 2 à 6. La figure FIG. 4.19 présente les TBC_{VIN} obtenus par les meilleures fusions d'un nombre n de critères, n allant de 2 à 6.

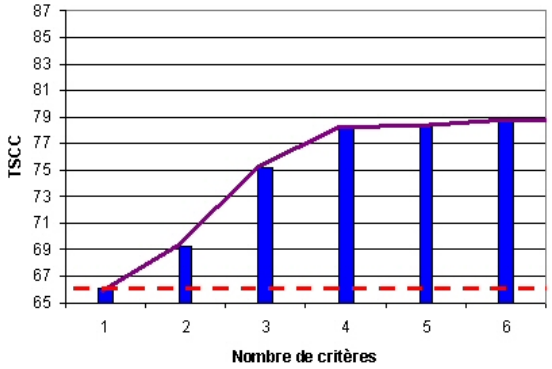


FIG. 4.19 – TBC_{VIN} de différentes combinaisons de critères sur la base globale (barres bleues) par rapport à l'optimal obtenu sans fusion (pointillés rouges).

Nous pouvons constater que, dès l'utilisation de deux critères, un gain de TBC_{VIN} apparaît même s'il n'est que de 3%. Le meilleur résultat est obtenu avec 6 critères pour un TBC_{VIN} de 78.7%, ce qui améliore nettement le meilleur résultat sans fusion qui est pour rappel de 66% pour le critère $LEV3$. Le temps nécessaire pour la détermination d'un vecteur de coefficients est d'environ 12 heures pour une combinaison de critères utilisant les paramètres que nous avons présentés, lorsque les calculs sont effectués sur un Pentium IV à 3Ghz et pour une implémentation sous ©Matlab [Houck *et al.* 1995b]. Afin d'obtenir une comparaison de résultats de segmentation plus proche de celle du critère VIN, nous pouvons utiliser comme critère amélioré la combinaison linéaire suivante :

$$CHA3 = 0.19LEV3 + 0.03ROS - 0.04ZEB - 0.99LEV4 - 0.18LEV1 - 1BOR \tag{4.4}$$

Nous obtenons dans ce cas $78,7\%$ de bonnes comparaisons sur la base de résultats testée.

Nous avons enfin opéré désormais la recherche des n coefficients pour la combinaison linéaire de n critères parmi les 6 à notre disposition (n allant de 2 à 6). L'algorithme génétique détermine alors à chaque étape (c'est à dire pour toutes valeurs prises par n) les n meilleurs critères parmi les 6 à fusionner.

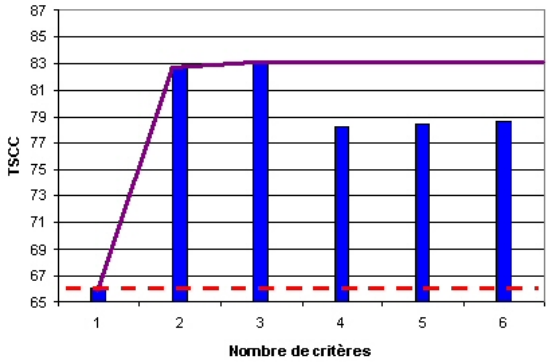


FIG. 4.20 – TBC_{VIN} de différentes combinaisons de critères déterminées par l'algorithme génétique sur la base globale (barres bleues) par rapport à l'optimal obtenu sans fusion (pointillés rouges).

La figure FIG. 4.20 présente les taux de bonne comparaison obtenus en opérant une sélection de critères. Toutes les combinaisons donnent une similarité de comparaison supérieure à $78,1\%$ et le nouveau critère par combinaison linéaire avec sélection est alors :

$$CHA4 = -0.68 * BOR - 0.01 * LEV4 + 0.003 * LEV3 \tag{4.5}$$

Le temps nécessaire pour la détermination et la sélection des coefficients est d'environ 14 heures pour une combinaison de critères, lorsque les calculs sont effectués sur un Pentium IV à 3Ghz et pour une implémentation sous ©Matlab [Houck *et al.* 1995b]. La combinaison et la sélection de plusieurs critères parmi les 6 permettent d'obtenir une meilleure similarité (TBC_{VIN} optimal de 83%) que celle fournie par un critère seul.

4.4.2.3 Fusion par apprentissage

Nous avons réalisé un apprentissage de la comparaison de deux résultats de segmentation. Cette approche est détaillée dans le chapitre précédent. Elle consiste à calculer des critères d'évaluation sur deux résultats de segmentation à comparer. Un SVM permet à partir de ces informations d'identifier parmi les deux résultats celui de meilleure qualité.

Dans un premier temps, nous nous intéressons au gain obtenu par la fusion de critères en utilisant une partie des résultats sur notre base globale en apprentissage et le reste comme base de test. Ainsi, nous présentons dans la figure FIG. 4.21 le TBC_{VIN} obtenu pour différentes proportions de la base globale utilisées en apprentissage, avec les éléments restants utilisés pour tester le résultat de la fusion.

La figure FIG. 4.22 présente le TBC_{VIN} obtenu selon le nombre de critères d'évaluation utilisés dans l'ensemble d'apprentissage.

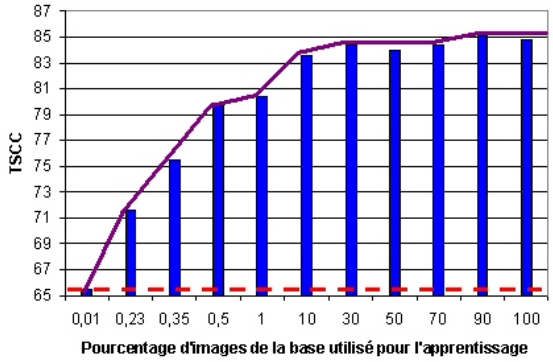


FIG. 4.21 – TBC_{VIN} de la fusion des 6 critères par apprentissage sur un pourcentage de la base globale (barres bleues) par rapport à l'optimal obtenu sans fusion (pointillés rouges).

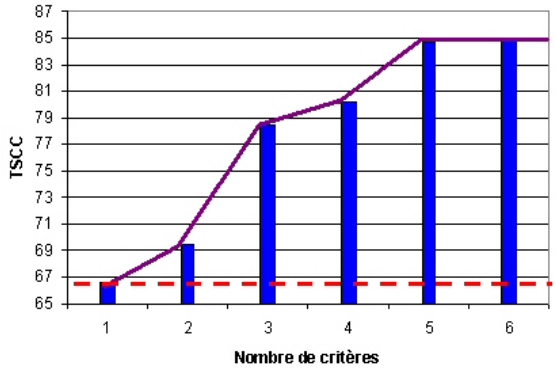


FIG. 4.22 – TBC_{VIN} de l'apprentissage de comparaisons de résultats de segmentation déterminé par un séparateur à vaste marge utilisant la base globale en apprentissage (barres bleues) par rapport à l'optimal obtenu sans fusion (pointillés rouges). Les critères fusionnés sont choisis par TBC_{VIN} décroissants.

Lorsque 5 critères d'évaluation sont utilisés dans la phase d'apprentissage, le taux de bonne reconnaissance sur l'ensemble des résultats de comparaison de la base est de 84.7%. L'apprentissage et la reconnaissance de classement des résultats d'évaluation de la base complète prennent en moyenne 70 minutes sur un Pentium IV à 3Ghz.

La figure FIG. 4.23 présente le TBC obtenu après une étape de sélection de critères d'évaluation. Les 6 critères ont été sélectionnés dans l'ordre suivant : BOR, ZEB, $LEV3$, $LEV4$, $LEV1$ puis enfin ROS. Lorsque 6 critères d'évaluation sont utilisés dans la phase d'apprentissage, le taux de bonne reconnaissance sur l'ensemble des résultats de comparaison de la base par ce critère $CHA6$ est de 85%. En n'utilisant que 4 critères après sélection, le TBC_{VIN} est de 84.5%.

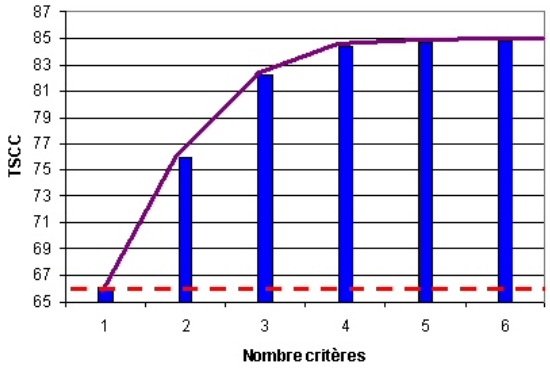


FIG. 4.23 – TBC_{VIN} de l'apprentissage de comparaisons de résultats de segmentation de n critères déterminé par un séparateur à vaste marge utilisant la base globale en apprentissage (barres bleues) par rapport à l'optimal obtenu sans fusion (pointillés rouges). Les critères fusionnés sont sélectionnés par le SVM.

4.4.2.4 Discussion

Le critère adaptatif permet d'atteindre un TBC_{VIN} de 73% soit un gain appréciable par rapport au meilleur critère de l'étude comparative ($LEV3$ avec un TBC_{VIN} de 66%). Dans un deuxième temps, nous avons fusionné les critères d'évaluation non supervisée par combinaison linéaire. Le résultat a donné un TBC_{VIN} de 79% en utilisant les 6 critères, puis nous avons sélectionné les critères à fusionner par l'algorithme génétique pour arriver à un TBC_{VIN} de 83% avec trois critères. Dans un troisième temps, nous avons réalisé un apprentissage de l'évaluation de résultats de segmentation. Nous avons obtenu un TBC_{VIN} de 85% en utilisant les 6 critères en apprentissage et 84% en utilisant les 4 critères sélectionnés.

Ces trois approches de fusion montrent qu'il est possible d'améliorer la pertinence d'un jugement réalisé par rapport à l'utilisation d'un unique critère d'évaluation. D'autres approches de fusion sont évidemment possibles. L'avantage des méthodes de fusion testées est que le jugement est basé sur un apprentissage, ce que fait habituellement un humain.

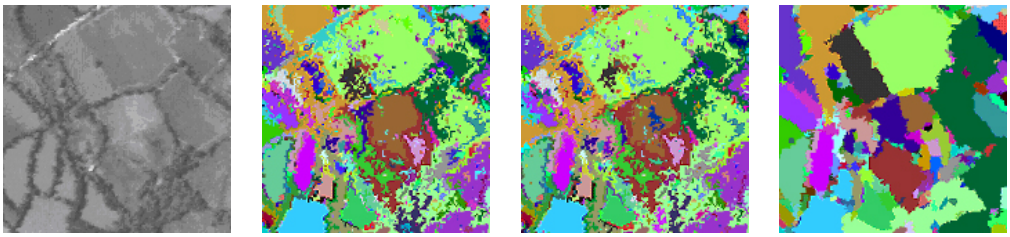
4.4.3 Applications des critères d'évaluation

Les critères d'évaluation non supervisée de résultats de segmentation ont un grand nombre d'applications. Nous présentons par la suite un certain nombre d'entre elles.

4.4.3.1 Comparaison de méthodes de segmentation

La comparaison de résultats de segmentation peut être réalisée à partir des critères d'évaluation explicités précédemment. Cette comparaison n'a évidemment de sens que

pour des résultats ayant le même niveau de précision. Le but est de comparer plusieurs méthodes de segmentation afin d'identifier leurs performances sur un type d'image ou des résultats issus de la même méthode avec différents paramètres. La figure FIG. 4.24 présente 3 résultats de segmentation d'une image par trois méthodes de segmentation. Ces résultats contiennent le même nombre de classes et sont coloriés par un procédé d'appariement afin de faciliter leur comparaison visuelle [Chabrier 2005].



(a) Image originale (b) FCM (c) PCM (d) EDISON

FIG. 4.24 – Comparaison de résultats de segmentation.

Le tableau TAB. 4.4 présente la valeur de quelques critères d'évaluation non supervisée de l'état de l'art calculés sur les trois résultats de segmentation de la figure FIG. 4.24. Les valeurs optimales (compte tenu du sens de variation des critères) sont présentées en gras. Comme le montre ce tableau, la grande majorité des critères préfèrent le résultat de segmentation EDISON qui contient effectivement des régions plus homogènes.

	FCM	PCM	EDISON
BOR	0.0222	0.0297	**0.0155**
ZEB	**0.6228**	0.6124	0.5428
LEV1	0.0901	0.0889	**0.1099**
LEV2	0.7258	0.7112	**0.9693**
LEV3	0.5202	0.5239	**0.5275**
ROS	0.6379	0.6328	**0.6973**

TAB. 4.4 – Valeur de 6 critères d'évaluation non supervisée des résultats de segmentation de la figure FIG. 4.24.

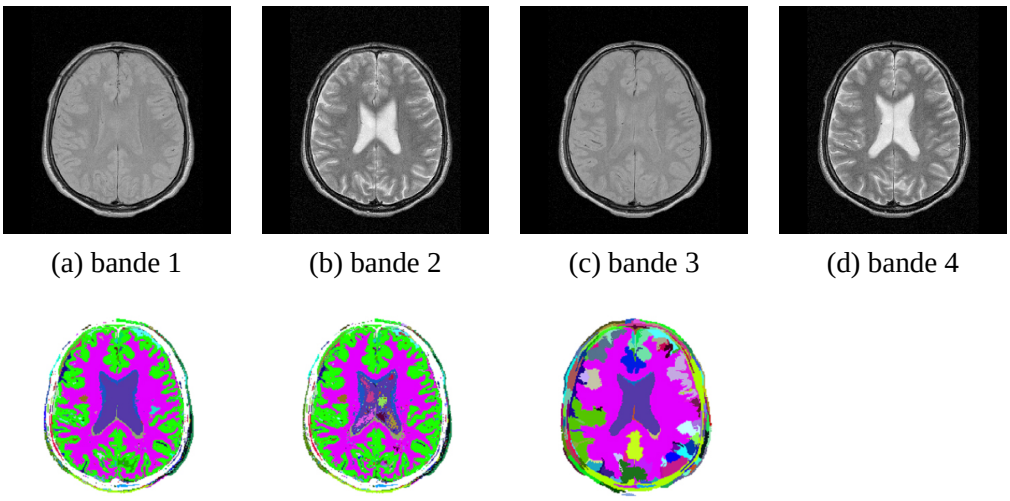
4.4.3.2 Evaluation de la qualité de résultats d'images multi-composantes

Il est possible de généraliser un critère d'évaluation afin de quantifier la qualité d'un résultat de segmentation d'une image multi-composantes [Chabrier *et al.* 2004a]. Pour ce faire, une méthode consiste à évaluer le résultat de segmentation en prenant en compte chaque composante. On obtient alors N valeurs du critère d'évaluation si l'image originale comporte N composantes. Il suffit alors de combiner ces valeurs pour obtenir

une valeur de qualité.

Nous avons réalisé une étude sur ce sujet en essayant plusieurs méthodes très simples de fusion telles que le minimum, le maximum ou la moyenne des N valeurs du critère [Chabrier *et al.* 2004a]. L'approche a été testée en générant une base d'images multi-composantes et en utilisant des vérités-terrain de la base de la figure FIG. 4.15. Chaque composante de l'image est associée à cette vérité terrain et comporte des textures et niveaux de gris différents des autres composantes. La moyenne des valeurs du critère a donné les meilleurs résultats après nos expérimentations.

La figure FIG. 4.25 présente 3 résultats de segmentation (de même nombre de classes) d'une image multi-composantes IRM par la méthode MLBG [Rosenberger et Chehdi 2003b].



(a) bande 1 (b) bande 2 (c) bande 3 (d) bande 4

(e) résultat 1 (f) résultat 2 (g) résultat 3

FIG. 4.25 – Comparaison de résultats de segmentation d'une image multi-composantes.

Le tableau TAB. 4.5 donne la valeur moyenne du critère ROS de chaque résultat de la figure FIG. 4.25. Le résultat 2 est considéré au sens de ce critère comme étant le meilleur.

Résultats de segmentation	Valeur du critère
résultat 1	0.66
résultat 2	**0.731**
résultat 3	0.649

TAB. 4.5 – Valeur du critère ROS moyen sur les résultats de segmentation de l'image multi-composantes de la figure FIG. 4.25.

4.4.3.3 Aide au paramétrage de méthodes de segmentation

Comme nous l'avons vu dans le chapitre 2 de ce manuscrit, une des caractéristiques d'un algorithme de traitement d'images et notamment de la segmentation, est le nombre important de paramètres à fixer. Un critère d'évaluation peut permettre de faciliter le paramétrage d'un algorithme de segmentation étant donnée une image à traiter. Pour illustrer ceci, nous avons réalisé des travaux visant à déterminer de façon automatique la valeur de deux paramètres de l'algorithme EDISON [Chabrier *et al.* 2005a]. Le tableau TAB. 4.6 présente l'évolution du nombre de régions dans le résultat de segmentation pour différentes valeurs des paramètres de l'algorithme EDISON. Nous avons distingué trois types de résultats de segmentation à partir du nombre de régions générées (cases grisées) :

- de 5 à 10 régions : faible précision,
- de 20 à 40 régions : précision moyenne,
- de 60 à 90 régions : grande précision.

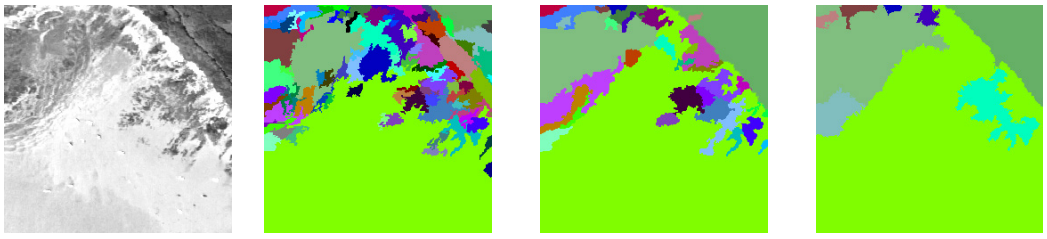
	3	5	7	9	11	13	15	17
2	86	85	82	90	86	82	73	75
5	70	**63**	60	50	46	45	38	37
8	41	41	36	28	35	25	17	24
11	**27**	21	22	20	18	12	13	**10**
14	16	15	13	14	13	10	8	7
17	5	7	8	11	9	5	8	5
20	5	4	4	7	6	3	6	6
23	3	2	2	8	4	5	4	2
26	1	6	1	1	2	2	2	2
29	1	2	1	1	1	1	1	3
31	3	1	1	1	1	1	1	1

TAB. 4.6 – Nombre de régions dans le résultat de segmentation pour chaque valeur des paramètres : spatial bandwidth (colonne) et range bandwidth (ligne) de l'algorithme EDISON.

Le tableau TAB. 4.7 présente l'évolution du critère de Zeboudj pour différentes valeurs des paramètres de l'algorithme EDISON. Ce critère d'évaluation a été utilisé compte tenu du caractère uniforme de l'image originale présentée dans la figure FIG. 4.26 (a). La valeur de ce critère permet de choisir parmi des résultats de segmentation de niveau de précision équivalent (ici, nous avons considéré le nombre de régions), le résultat de meilleure qualité au sens du critère de Zeboudj. Ainsi, la figure FIG. 4.26 présente les résultats optimaux au sens de ce critère pour les différents niveaux de précision associés aux valeurs des deux paramètres à régler.

	3	5	7	9	11	13	15	17
2	0,53	0,54	0,54	0,53	0,53	0,53	0,55	0,54
5	0,56	**0,57**	0,56	0,54	0,54	0,53	0,55	0,54
8	0,59	0,55	0,58	0,54	0,52	0,5	0,53	0,54
11	**0,59**	0,52	0,54	0,53	0,51	0,51	0,52	**0.51**
14	0,49	0,49	0,48	0,49	0,49	0,47	0,44	0,44
17	0,01	0,09	0,35	0,47	0,49	0,44	0,44	0,42
20	0,01	0,01	0	0,35	0,35	0,14	0,41	0,39
23	0	0	0	0,30	0,02	0,02	0,07	0,03
26	0	0	0	0	0,03	0,03	0,03	0
29	0	0	0	0	0	0	0	0
31	0	0	0	0	0	0	0	0

TAB. 4.7 – Evolution du critère de Zeboudj pour chaque valeur des paramètres : spatial bandwidth (colonne) et range bandwidth (ligne) de l'algorithme EDISON.

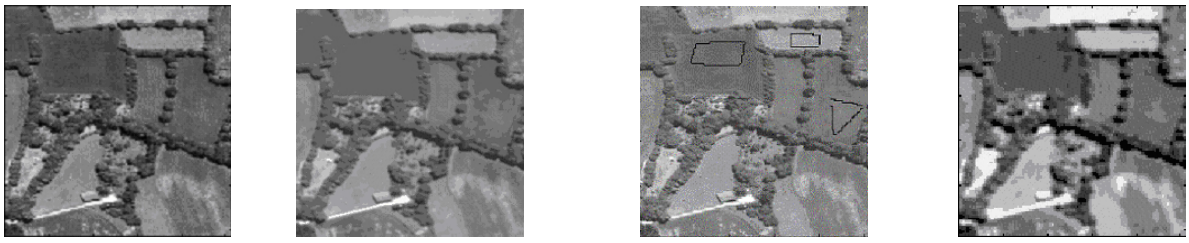


(a) Image originale (b) précision élevée (c) précision moyenne (d) précision basse

FIG. 4.26 – Résultats de segmentation obtenus après détermination du paramétrage de la méthode de segmentation.

4.4.3.4 Segmentation par optimisation de critères d'évaluation

L'optimisation d'un critère d'évaluation permet la définition de nouvelles méthodes de segmentation. La segmentation en régions consiste à chercher pour chaque pixel de l'image une étiquette de région ou de classe suivant le cas. Un critère d'évaluation permet de quantifier dans quelle mesure une solution potentielle à ce problème est pertinente. Nous avons adopté cette approche en proposant en 2002 deux méthodes de segmentation où un algorithme génétique était utilisé pour cette recherche [Rosenberger et Chehdi 2002]. La première méthode consiste à utiliser un critère d'évaluation non supervisée (ici le critère ROS) afin d'obtenir une méthode de segmentation automatique. La seconde consiste à ajouter une contrainte sur le niveau de précision souhaité pour obtenir une méthode semi-supervisée (d'après notre nomenclature). Dans ce cas, l'utilisateur sélectionne des zones de l'image signifiant que les pixels d'une zone doivent appartenir à une même région ou classe dans le résultat de segmentation final. Le résultat final est un compromis entre la qualité du résultat de segmentation et le respect de la contrainte sur la précision du résultat. La figure FIG. 4.27 présente les résultats de ces deux méthodes sur une bande d'une image CASI.

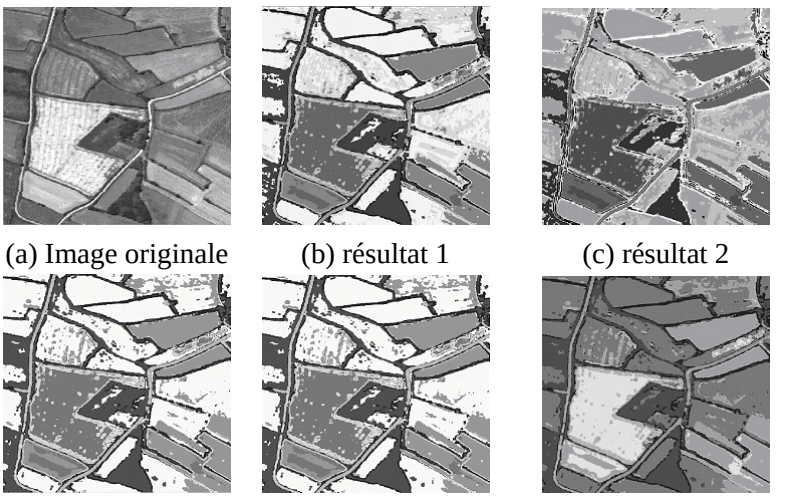


(a) Image originale (b) résultat non supervisé (c) définition de la précision (d) résultat supervisé

FIG. 4.27 – Segmentation par optimisation de critères d'évaluation.

4.4.3.5 Fusion de résultats de segmentation

La fusion de résultats de segmentation a de nombreux intérêts. Elle permet d'une part de combiner des résultats de différentes méthodes de segmentation ou d'une même méthode de segmentation paramétrée différemment. D'autre part, elle est indispensable lors de la segmentation scalaire d'une image multi-composantes (traitement en parallèle de chaque composante). Un critère d'évaluation est utilisé dans ce cas comme un indice de confiance ou de pertinence d'un résultat lors de la fusion. Il est possible d'exploiter un algorithme génétique pour déterminer automatiquement les informations à extraire des résultats de segmentation pour la fusion. La figure FIG. 4.28 présente le résultat de fusion de 4 résultats de segmentation d'une bande d'une image CASI (Compact Airborne Spectrographic Imager) en utilisant le critère ROS.

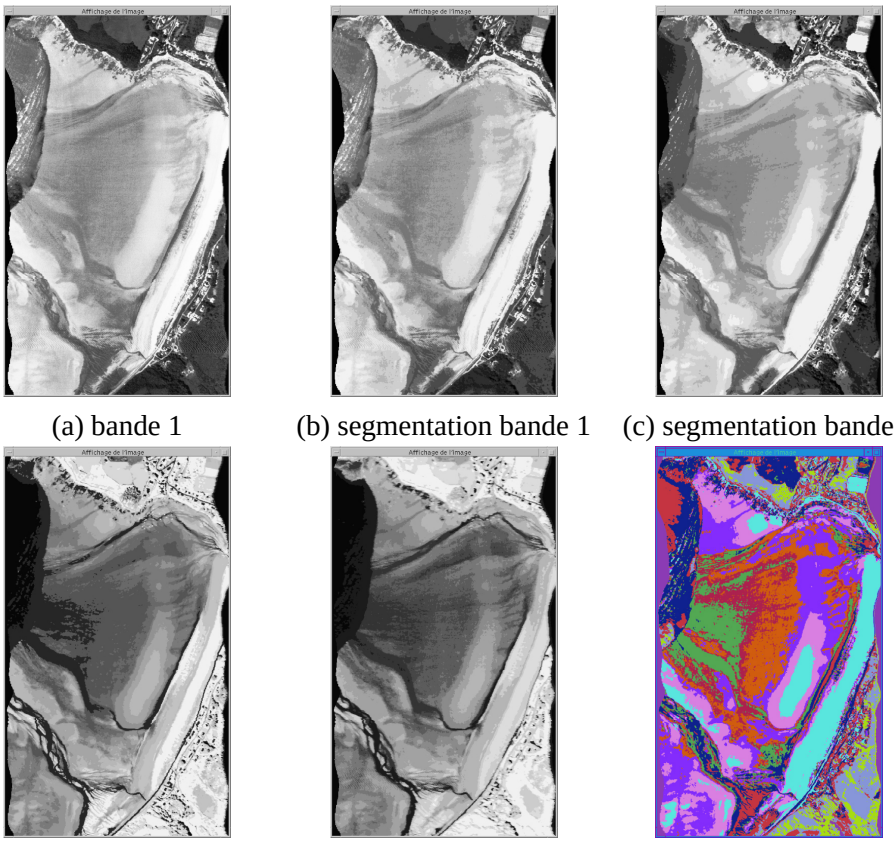


(a) Image originale (b) résultat 1 (c) résultat 2

(d) résultat 3 (e) résultat 4 (d) résultat de la fusion

FIG. 4.28 – Fusion de résultats de segmentation.

Nous présentons un résultat obtenu sur une image acquise par le capteur CASI de taille 1024×512 pixels avec 10 bandes spectrales (cf. FIG. 4.29). Chaque bande spectrale a été segmentée en utilisant une méthode de multi-seuillage [Kermad *et al.* 2000]. Le critère d'évaluation utilisé est la moyenne des valeurs du critère ROS sur chaque bande. Le résultat de segmentation obtenu est satisfaisant visuellement et prend en compte les informations complémentaires dans chaque bande. Le tableau TAB. 4.8 présente l'évolution du critère ROS dans l'algorithme génétique. On constate tout d'abord, une évolution assez nette de la qualité du résultat de la fusion (au sens de la valeur du critère). Enfin, l'algorithme a convergé vers un optimum de ce critère.



(a) bande 1 (b) segmentation bande 1 (c) segmentation bande 6

(d) segmentation bande 7 (e) segmentation bande 9 (f) résultat de la fusion

FIG. 4.29 – Fusion de résultats de segmentation dans le cas multi-composantes.

4.5 Conclusion et perspectives

Dans ce chapitre, nous avons présenté un ensemble de travaux sur l'évaluation de la segmentation d'images.

Information	Valeur du critère
critère d'évaluation moyen de la population initiale	0.903
meilleur critère d'évaluation de la population initiale	0.913
moins bon critère d'évaluation de la population initiale	0.881
écart type des critères d'évaluation de la population initiale	0.008
critère d'évaluation moyen de la population finale	0.979
meilleur critère d'évaluation de la population finale	0.979
moins bon critère d'évaluation de la population finale	0.979
écart type des critères d'évaluation de la population finale	6.02e-08

TAB. 4.8 – Quelques statistiques du critère d'évaluation des populations initiale et finale lors de la segmentation de l'image multi-composante de la figure FIG. 4.29.

Nous avons contribué à l'analyse des performances des critères d'évaluation de la qualité d'un résultat de segmentation. Ce travail a permis de mieux cerner les critères les plus pertinents de l'état de l'art. A l'heure actuelle, le critère le plus utilisé est sans doute le critère de Borsotti BOR pour des résultats de segmentation en régions. La raison majeure d'un tel choix est sa simplicité de calcul (comme le PSNR en compression). Néanmoins, comme nous l'avons vu, ce critère est loin d'être efficace surtout lorsque l'image contient un peu de texture. Les études comparatives respectant un protocole rigoureux sont nécessaires afin que la communauté scientifique utilise des critères pertinents.

Nous avons mis en évidence l'intérêt de fusionner plusieurs critères d'évaluation pour obtenir un jugement plus fiable. Cette approche est sans doute plus efficace que la conception d'un nouveau critère d'évaluation qui reste une tâche très difficile. Nous avons montré enfin les différentes motivations pour disposer d'un critère d'évaluation performant. De nombreuses applications sont associées à l'utilisation de ces critères.

Les perspectives sur l'évaluation de la segmentation sont nombreuses.

L'évaluation de la segmentation est actuellement un sujet intéressant de plus en plus de chercheurs. Il est possible d'attendre d'importants progrès dans ce domaine si des bases d'images, de résultats de segmentation et des implementations de critères sont disponibles pour la communauté. J'ai mis sur ma page personnelle ce type d'informations afin que des chercheurs puissent télécharger un critère d'évaluation pour valider leur algorithme de segmentation ou comparer l'efficacité d'un critère d'évaluation. Une large diffusion de l'efficacité des critères d'évaluation est souhaitable.

La comparaison de critères d'évaluation peut nécessiter des bases de résultats de segmentation générés manuellement par des experts. La base de Berkeley [Martin *et al.* 2001] souvent utilisée, pose de nombreux problèmes de définition. Le niveau de précision des résultats dans cette base sont très différents d'un expert à

un autre. La génération d'une base significative de résultats de segmentation à partir d'images naturelles représentatives serait très utile à la communauté. Ceci suppose une réflexion sur la formalisation du niveau de précision d'un résultat de segmentation. Une méthode permettant de garantir un niveau de précision identique pour des résultats de segmentation générés manuellement doit être mise au point.

La plupart des critères d'évaluation non supervisée ne prennent pas en compte la notion de texture dans leur calcul. Même si comme nous l'avons montré, la disparité inter-classes est plus discriminante que l'intra-classe, l'utilisation de paramètres de texture me semble indispensable.

Enfin, l'utilisation de procédés de fusion plus élaborés peut être explorée afin de tirer parti des informations complémentaires apportées par les différents critères d'évaluation.

Conclusion et projet de recherche

Le traitement d'images a atteint une bonne maturité après plus de trente ans de recherche. Après l'essor du développement d'algorithmes de traitement d'images, il devient nécessaire d'évaluer leurs performances afin d'obtenir un retour d'expérience sur l'existant. L'objectif est d'identifier les verrous scientifiques et donc d'orienter les recherches futures en traitement d'images.

L'évaluation est un sujet intéressant et très important. L'essor d'Internet a permis l'échange de publications, d'algorithmes et de données test entre tous les chercheurs dans ce domaine. Plusieurs cadres méthodologiques d'évaluation d'algorithmes ont été proposés dans la littérature, ils ont fait l'objet du chapitre 2 de ce manuscrit. L'approche privilégiée par beaucoup de conférences, programmes de recherche (comme Technovision) et projets est l'approche dite d'évaluation par diagnostic. Cette approche permet un jugement fiable au prix de beaucoup d'efforts pour la conception des bases de données test et des vérités terrain associées.

Nous avons mis en évidence l'importance des critères de performance dans ce cadre. Les critères d'évaluation supervisée exploitent une information supplémentaire pour réaliser un jugement. Celle-ci permet en général un accroissement de la pertinence de l'évaluation au prix d'un effort supplémentaire pour la concevoir. Les critères d'évaluation non supervisée sont complètement automatisables, ce qui est un avantage indéniable. En général, ils procurent néanmoins un jugement moins fiable. Les critères hybrides tentent de par leur conception d'exploiter cette complémentarité évidente de ces deux dernières catégories.

Nous avons présenté dans le chapitre 3 des travaux qui ont été réalisés sur l'évaluation de la compression d'images ultrasonores. Le contexte médical de l'étude a montré l'importance de la subjectivité du jugement d'autant plus que des notions de perception visuelle du résultat intervenaient. Un cadre méthodologique a été appliqué consistant à étudier des critères d'évaluation non supervisée en se basant sur le jugement d'expert. Cela a permis la définition de critères hybrides ayant un comportement plus proche de

l'expert.

Le chapitre 4 présente les activités réalisées dans le cadre de l'évaluation de la segmentation. Nous nous sommes proposés dans un premier temps de formaliser l'étape de segmentation afin d'identifer clairement ses objectifs. Le cadre méthodologique utilisé en compression a été appliqué de façon similaire. Il faut souligner l'effort important apporté sur la quantité et la représentativité des images synthétiques utilisées lors de l'étude comparative. La proposition de nouveaux critères hybrides s'est cette fois basée sur des données synthétiques, étant donnée la richesse des informations à disposition. Nous avons mis en évidence les nombreuses applications découlant de l'utilisation de ces critères.

Quel est mon projet de recherche ? J'ai identifié deux axes majeurs de recherche.

Jusqu'à présent, j'ai travaillé essentiellement sur des traitements bas niveau dans la chaîne d'analyse d'images. Mon objectif est de me focaliser plus particulièrement sur des algorithmes de plus haut niveau.

Le premier axe de recherche concerne ainsi le développement de systèmes d'interprétation d'images. L'objectif de ce type de traitement est l'identification et la localisation d'objets d'intérêt dans une image. Un être humain est capable de localiser un objet particulier dans son environnement. Il utilise pour cela une base de connaissances qui s'étoffe au fil des années. L'identification d'un objet est réalisée en effectuant une mise en correspondance entre des parties d'un objet inconnu et celles des objets appris.

Il est à mon avis possible, étant donnés les grands progrès réalisés en reconnaissance d'objets [Everingham *et al.* 2006], de développer un système générique réalisant cette tâche. Le terme générique signifie qu'il peut être utilisé pour différentes applications en ne modifiant que les connaissances à disposition pour réaliser le traitement. Classiquement, afin de localiser un objet, un traiteur d'images applique un certain nombre d'opérateurs tels que les pré-traitements ou la segmentation. Étant donnée la complexité de la tâche de segmentation, l'information extraite (localisation de régions ou de contours) ne facilite pas forcément l'interprétation de l'image. Il est à mon avis beaucoup plus pertinent de travailler directement sur l'image brute d'acquisition et d'utiliser des techniques de reconnaissance de formes. Il est possible de rendre ces techniques plus robustes aux différents artefacts pouvant altérer la qualité de l'image (lumière, bruit, occultation...). Les approches locales consistant à reconnaître un objet à partir de plusieurs de ses sous-parties permettent d'obtenir de bons résultats.

Je co-encadre une thèse sur ce sujet (thèse d'Anant Choksuriwong). Les objectifs de ces travaux sont de developer des algorithmes de détection et de reconnaissance d'objets dans une image. Les applications envisagées sont de de contribuer notamment à doter un robot mobile d'une meilleure capacité de perception de l'environnement (qui est une

thématique forte de l'équipe Robotique du laboratoire LVR).

Le second axe de recherche est de continuer à travailler sur l'évaluation d'algorithmes de traitement d'images. Le cadre méthodologique d'évaluation appliqué en compression et segmentation peut être généralisé à tout traitement. Une condition préalable est toutefois la définition des objectifs de ce traitement. L'évaluation de pré-traitements d'une image tels que le filtrage et la restauration est souvent réalisée dans un cadre supervisé (comapraison du résultat du traitement d'une image altérée artificiellement). Il serait intéressant de développer des critères d'évaluation non supervisée dans ce contexte.

Les études de critères d'évaluation sont assez rares dans la littérature mais pourtant fondamentales pour que la communauté scientifique utilise les bons outils d'évaluation pour valider leurs algorithmes. Différents travaux peuvent être réalisés sur l'analyse d'algorithmes de traitement d'images. Il est possible par exemple d'étudier à l'aide de critères d'évaluation pertinents, l'impact du type d'image originale ou de paramètres d'un algorithme sur sa performance. Ce type d'étude permettra une meilleure analyse de ces algorithmes et facilitera leurs applications en conditions réelles dans l'industrie.

Comme je l'ai mentionné précédemment, les traitements haut-niveau m'intéressent plus particulièrement. Il est primordial de quantifier dans quelle mesure la tâche d'interprétation d'une image est correctement réalisée. La majeure partie des systèmes d'interprétation a pour finalité l'identification d'un objet dans l'image (biométrie, télé-détection...), la mise en évidence d'une situation (vidéosurveillance, vision embarquée en automobile...) ou la quantification d'un événement (contrôle qualité). Plusieurs problématiques en découlent. La détection est-elle correctement réalisée (localisation, reconnaissance, précision) ? Y a t-il des informations non détectées par l'algorithme ? Cette caractérisation peut être encore une fois être réalisée en mode supervisé (en exploitant une vérité terrain) ou en mode non supervisé. Une thèse va débuter en septembre 2006 sur ce sujet pour initier ces travaux. Dans le cas de l'interprétation, l'objectif est clairement identifié. La génération d'une vérité terrain sur des images représentatives ne pose pas de problème majeur. De nombreuses données seront disponibles au travers de l'action Technovision avec les vérités terrain associées. Elles permettront d'évaluer la performance d'algorithmes de traitement d'images mais aussi de grands progrès dans la définition de métriques fiables pour leurs validations.

Ces deux axes de recherche correspondent à ceux que j'ai définis à partir de 2003 lors de la création du groupe Interprétation au sein du LVR. En ce qui concerne le premier axe, le nombre d'applications potentielles de l'interprétation d'images est tellement grand qu'il me parait opportun de contribuer au développement de tels algorithmes. Le second axe me permettra d'affirmer une spécificité dans le contexte national et international. La définition d'outils d'évaluation est indispensable pour garantir des progrès du domaine de la conception d'algorithmes de traitement d'images. Ceci constitue néanmoins un défi majeur. En effet, une métrique d'évaluation pour pouvoir être adoptée par la communauté

doit être validée avec beaucoup de rigueur. A l'heure actuelle, la plupart des chercheurs utilisent une métrique (par exemple le PSNR) parce qu'elle est couramment exploitée dans la littérature même si d'autres sont plus pertinentes. Outre la difficulté de conception de ces critères d'évaluation, se dégage donc, la nécessité d'une communication importante de leur efficacité. C'est bien aux travers de ces deux axes que je souhaite m'employer dans mes activités futures de recherche et d'encadrement et donc continuer à apporter ma contribution à l'évaluation d'algorithmes de traitements d'images.

Liste des tableaux

1.1 Etat civil 7
1.2 Parcours 7
1.3 Activités pédagogiques et administratives 8
1.4 Activités de recherche 8
1.5 Enseignements dispensés à l'ENSI de Bourges 9
1.6 Financements obtenus 18

2.1 Tableau récapitulatif des méthodes d'évaluation de performance 41

3.1 Les 21 critères objectifs d'évaluation de résultats de compression d'images étudiés. 58

4.1 Classement moyen de chaque résultat de segmentation par les différents critères, 1 indiquant le meilleur résultat. 89
4.2 Présentation du taux de bonne comparaison des critères d'évaluation supervisée de résultats de segmentation en régions par rapport au classement d'experts. 92
4.3 TBC des critères d'évaluation non supervisée par rapport au critère VIN sur différentes sous-bases d'images. 95
4.4 Valeur de 6 critères d'évaluation non supervisée des résultats de segmentation de la figure FIG. 4.24. 102
4.5 Valeur du critère ROS moyen sur les résultats de segmentation de l'image multi-composantes de la figure FIG. 4.25. 103
4.6 Nombre de régions dans le résultat de segmentation pour chaque valeur des paramètres : spatial bandwidth (colonne) et range bandwidth (ligne) de l'algorithme EDISON. 104
4.7 Evolution du critère de Zeboudj pour chaque valeur des paramètres : spatial bandwidth (colonne) et range bandwidth (ligne) de l'algorithme EDISON. 105
4.8 Quelques statistiques du critère d'évaluation des populations initiale et finale lors de la segmentation de l'image multi-composante de la figure FIG. 4.29. 108

Table des figures

1.1 Exemple de projet étudiant encadré : Le robot qui joue au billard 10
1.2 Description d'un objet par des invariants locaux. 14
1.3 Exemple d'un objet 3D modélisé. 14

2.1 Un algorithme de traitement d'images 30
2.2 L'exemple de la compression d'images (compression JPEG avec un taux de compression de 20%) : à gauche, l'image originale et à droite le résultat de compression. 32
2.3 Evaluation par diagnostic 33
2.4 Deux exemples de vérité terrain utilisée dans la compétition VOC ici sur la détection d'un vélo dans une image 34
2.5 Evaluation supervisée 36
2.6 Exemples d'images synthétiques 37
2.7 Exemple d'une image benchmark 38
2.8 Exemples de vérités-terrain réalisées manuellement pour la segmentation d'images naturelles par trois experts. 38
2.9 Comparaison de résultats de compression d'images ultrasonores 39
2.10 Evaluation non supervisée 40

3.1 La chaîne de télé-échographie OTELO. 44
3.2 Le système OTELO : (a) le poste expert, (b) la sonde fictive et (c) le poste patient. 45
3.3 Mise en évidence de la distortion engendrée par la compression d'une image. 46
3.4 Différentes méthodes d'évaluation de la qualité d'un résultat de compression d'une image dans la littérature. 48
3.5 Evaluation subjective à un stimulus. 53
3.6 Evaluation subjective comparative. 55
3.7 Classement des résultats de compression : l'image B est dans ce cas considérée comme étant la plus proche de l'image originale. 56
3.8 Principe de définition de la comparaison. 59
3.9 Taux de Bonne Comparaison TBC de chaque critère avec l'évaluation des experts (classés dans l'ordre décroissant). 60

3.10 Mise en évidence du pouvoir discriminant des critères objectifs (le symbole * signifie que le résultat de compression A est de meilleure qualité par rapport au résultat B selon les experts). 61
3.11 Fusion de critères d'évaluation objectifs. 62
3.12 Généralisation de la fusion par combinaison linéaire. 63
3.13 Taux de Bonne Comparaison TBC par fusion avec la combinaison linéaire sans sélection. 64
3.14 Taux de Bonne Comparaison TBC par fusion avec la combinaison linéaire avec sélection. 66
3.15 Mise en évidence du pouvoir discriminant du critère $DEL2$. 67
3.16 Fusion de critères objectifs par un SVM. 68
3.17 Généralisation de la fusion par un SVM. 69
3.18 Sélection des variables pour la fusion avec un SVM. 69

4.1 La chaîne de traitement d'une image (images du ©LASTI). 74
4.2 Les différentes catégories de méthodes de segmentation. 76
4.3 La texture sable. 77
4.4 Différents résultats de segmentation d'une image. $a)$ image originale, $b)$ segmentation en 7 régions, $c)$ segmentation en 3 classes, $d)$ image originale, $e)$ segmentation en 38 régions, $f)$ segmentation en 3 classes. 78
4.5 Évaluation supervisée d'un résultat de segmentation à l'aide d'un tracé expert. 79
4.6 Évaluation non supervisée d'un résultat de segmentation à l'aide de calculs statistiques. 81
4.7 Découpage en blocs du résultat de segmentation à évaluer. 84
4.8 Plusieurs dégradations synthétiques d'une vérité terrain en contour. . . . 86
4.9 Evolution de trois critères d'évaluation de détection de contours face à l'ajout de bruit impulsionnel sur la vérité terrain (l'abscisse des courbes correspondent aux différents résultats synthétiques de segmentation et l'ordonnée à la valeur de chaque critère). 87
4.10 Différents résultats de segmentation extraits manuellement. 87
4.11 Résultats de segmentation utilisés pour la comparaison de critères d'évaluation supervisée. 88
4.12 Interface du test pour la comparaison de résultats de segmentation. 90
4.13 Plusieurs dégradations synthétiques d'une vérité terrain en régions. 91
4.14 Evolution de trois critères d'évaluation région face à la sous-segmentation (l'abscisse des courbes correspondent aux différents résultats synthétiques de segmentation et l'ordonnée à la valeur de chaque critère). 91
4.15 Vérités-terrain utilisées pour la création des bases d'images synthétiques de 2 à 15 classes. 93
4.16 Exemple d'une image comportant 5 classes et ses résultats de segmentation avec couleurs appariées. 94
4.17 Principe du critère d'évaluation adaptatif. 96

4.18 TBC_{VIN} de la fusion des 6 critères par l'algorithme génétique sur un pourcentage de la base globale (barres en bleu) par rapport à l'optimal obtenu sans fusion (pointillés rouges). 97
4.19 TBC_{VIN} de différentes combinaisons de critères sur la base globale (barres bleues) par rapport à l'optimal obtenu sans fusion (pointillés rouges). 98
4.20 TBC_{VIN} de différentes combinaisons de critères déterminées par l'algorithme génétique sur la base globale (barres bleues) par rapport à l'optimal obtenu sans fusion (pointillés rouges). 99
4.21 TBC_{VIN} de la fusion des 6 critères par apprentissage sur un pourcentage de la base globale (barres bleues) par rapport à l'optimal obtenu sans fusion (pointillés rouges). 100
4.22 TBC_{VIN} de l'apprentissage de comparaisons de résultats de segmentation déterminé par un séparateur à vaste marge utilisant la base globale en apprentissage (barres bleues) par rapport à l'optimal obtenu sans fusion (pointillés rouges). Les critères fusionnés sont choisis par TBC_{VIN} décroissants. 100
4.23 TBC_{VIN} de l'apprentissage de comparaisons de résultats de segmentation de n critères déterminé par un séparateur à vaste marge utilisant la base globale en apprentissage (barres bleues) par rapport à l'optimal obtenu sans fusion (pointillés rouges). Les critères fusionnés sont sélectionnés par le SVM. 101
4.24 Comparaison de résultats de segmentation. 102
4.25 Comparaison de résultats de segmentation d'une image multi-composantes. 103
4.26 Résultats de segmentation obtenus après détermination du paramétrage de la méthode de segmentation. 105
4.27 Segmentation par optimisation de critères d'évaluation. 106
4.28 Fusion de résultats de segmentation. 106
4.29 Fusion de résultats de segmentation dans le cas multi-composantes. . . . 107

Article

EURASIP Journal on Applied Signal Processing

Special Issue on Performance Evaluation in Image Processing

Article

IEEE Transactions on Medical Imaging
Special Issue on Validation in Medical Image Processing

Bibliographie

[Aiazzi *et al.* 1997] B. Aiazzi, L. Alparone, S. Baronti, et G. Borri. *Multiresolution adaptive speckle filtering : a comparison of algorithms*. In *IEEE International Conference on Geoscience and Remote Sensing (IGARSS)*, volume 2, pages 1054–1056, 1997.

[Andrey 1999] Philippe Andrey. *Selectionist Relaxation : Genetic Algorithms Applied to Image Segmentation*. Image and Vision Computing, 17 : 175–187, 1999.

[Androustos *et al.* 1998] D. Androustos, K.N. Plataniotis, et A.N. Venetsanopoulos. *Distances Measures for Color Image Retrieval*. IEEE ICIP International Conference on Image Processing, 1998.

[Argenti et Alparone 2002] F. Argenti et L. Alparone. *Speckle removal from SAR images in the undecimated wavelet domain*. IEEE Transactions on Geoscience and Remote Sensing, 40 : 2363–2374, 2002.

[Avcibas *et al.* 2002] Ismail Avcibas, Bülent Sankur, et Khalid Sayood. *Statistical evaluation of image quality measures*. Journal of Electronic imaging, 11(2) : 206–223, avril 2002.

[Azpiroz-Leehan *et al.* 2004] J. Azpiroz-Leehan, R. Leder, et J.-F. Lerallut. *Quantitative and qualitative evaluation of filter characteristics for wavelet packet compression of MR images*. In *26th Annual International Conference of the Engineering in Medicine and Biology Society*, pages 1537 – 1540, 2004.

[Baddeley 1992] A. J. Baddeley. *An Error metric for binary images*. Robust Computer Vision, pages 59–78, 1992.

[Bailey et Srinath 1996] R. Bailey et M. Srinath. *Orthogonal Moment Features for Use With Parametric and Non-Parametric Classifiers*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 18(4) : 389–399, 1996.

[Basseville 1989] M. Basseville. *Distance measures for signal processing and pattern recognition*. Signal processing, 18 (4) : 349–369, 1989.

[Baudin *et al.* 1996] Olivier Baudin, Atilla Baskurt, Thieery Moll, Rémy Prost, Didier Revel, Fenno Ottes, Mohammed Khamadja, et Michel Amiel. *ROC assessment of compressed wrist radiographs*. European Journal of Radiology, 22 : 228–231, 1996.

[Beauchemin *et al.* 1998] M. Beauchemin, KP. B. Thomson, et G. Edwards. *On The Hausdorff Distance Used For The Evaluation Of Segmentation Results*. Canadian Journal of Remote Sensing, 24(1) : 3–8, 1998.

[Beaurepaire 1997] Patricia Beaurepaire. *Compression d'Images appliquée aux Angiographies Cardiaques : Aspects algorithmiques, Evaluation de la qualité diagnostique*. Thèse de Doctorat, Université de Technologie de Compiègne, novembre 1997.

[Bebis et Miller 2006] G. Zehang Sun Bebis et R. Miller. *On-Road Vehicle Detection : A Review*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 28(5) : 694–711, 2006.

[Beegan *et al.* 2002] A.P. Beegan, L.R. Iyer, A.E. Bell, V.R. Maher, et M.A. Ross. *Design and evaluation of perceptual masks for wavelet image compression*. In *Proceedings of 2002 IEEE 10th Digital Signal Processing Workshop*, pages 88–93, 2002.

[Beichel *et al.* 2005] R. Beichel, H. Bischof, F. Leberl, et M. Sonka. *Robust active appearance models and their application to medical image analysis*. IEEE Transactions on Medical Imaging, 24 : 1151–1169, 2005.

[Belaroussi *et al.* 2002] B. Belaroussi, C. Odet, et H. Benoit-Cattin. *Scalable discrepancy measures for segmentation evaluation*. In *In IEEE International Conference on Image Processing (ICIP)*, volume 1, pages 785–788, 2002.

[Bezdeck 1981] J. Bezdeck. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press, 1981.

[Bhandakar *et al.* 1994] S.M. Bhandakar, Y. Zhang, et W.D. Potter. *An edge detection technique using genetic algorithm-based optimisation*. Pattern Recognition, 9(27) : 1159–1180, 1994.

[Blanc *et al.* 1999] D. Blanc, A. Gourdon, P. Marché, Y. Parmantier, Ph. Poignet, G. Poisson, et P. Vieyres. *Robot à Trois Degrés de Liberté et à un Point Fixe*. , brevet français N 99 03 736, mars 1999.

[Borsotti *et al.* 1998] M. Borsotti, P. Campadelli, et R. Schettini. *Quantitative evaluation of color image segmentation results*. Pattern Recognition Letters, 19 : 741–747, 1998.

[Bougeniere *et al.* 2005] G. Bougeniere, P. Moulon, C. Rosenberger, et W. Smari. *On the determination of 3D Trajectory of Moving Targets by Stereovision*. In *IEEE International Conference on Computers, Communications, and Signal Processing (CCSP)*, 2005.

[Burgul *et al.* 2000] R. Burgul, F.J. Gilbert, et P.E. Undrill. *Methods of measurement of image quality in teleultrasound*. British Journal of Radiology, 73(876) : 1306–1312, 2000.

[Callet et Barba 2001] Patrick Le Callet et Dominique Barba. *Critère de qualité visuelle d'images couleur - application à l'évaluation réaliste de codeurs d'images*. GRETSI, septembre 2001.

[Cane 1997] B. Mc Cane. *On the Evaluation of Image Segmentation Algorithms*. DICTA'97 and IVCN ' Z ' 97, pages 455–459, october 1997.

[Canu *et al.* 2003] S. Canu, Y. Grandvalet, et A. Rakotomamonjy. *SVM and Kernel Methods Matlab Toolbox*. , Perception Systèmes et Information and INSA de Rouen and Rouen and France, 2003.

[Cariou *et al.* 1999] C. Cariou, J.M. Ogier, S. Adam, R. Mullot, J. Gardes, et Y. Lecourtier. *A Multi-Scale and Multi-Orientation Recognition Technique Applied to Document Interpretation : Application to French Telephone Network Maps*. International Journal Pattern Recognition Artificial Intelligence : special number on invariants, 1 : 1201–1218, 1999.

[Carnec 2004] Mathieu Carnec. *Critères de qualité d'images couleur avec référence réduite perceptuelle générique*. Thèse de Doctorat, Ecole Polytechnique de l'Université de Nantes, juillet 2004.

[Cha et Srihari 2002] Sung-Hyuk Cha et Sargur N. Srihari. *On measuring the distance between hostograms*. Pattern Recognition, 35 : 1355–1370, 2002.

[Chabrier 2005] S. Chabrier. *Contribution à l'évaluation de performances de la segmentation d'images*. Thèse de Doctorat, Université d'Orléans, 2005.

[Chabrier *et al.* 2004a] S. Chabrier, B. Emile, H. Laurent, C. Rosenberger, et P. Marché. *Unsupervised Evaluation of Image Segmentation : Application to multi-spectral images*. IEEE International Conference on Pattern Recognition, 3 : 576–579, 2004.

[Chabrier *et al.* 2004b] S. Chabrier, H. Laurent, B. Emile, C. Rosenberger, et P. Marché. *A comparative study of supervised evaluation criteria for image segmentation*. EUSIPCO, pages 1143–1146, 2004.

[Chabrier *et al.* 2005a] S Chabrier, H. Laurent, et B. Emile. *Performance evaluation of image segmentation. Application to parameters fitting*. EUSIPCO, 2005.

[Chabrier *et al.* 2005b] S. Chabrier, H. Laurent, C. Rosenberger, et A. Rakotomamonjy. *Fusion de critères pour l'évaluation de résultats de segmentation d'images*. GRETSI, pages CD–ROM, septembre 2005.

[Chabrier *et al.* 2005c] S. Chabrier, C. Rosenberger, et B. Emile. *Segmentation Evaluation by Fusion with a Genetic Algorithm*. EUSIPCO, 2005.

[Chabrier *et al.* 2006a] S. Chabrier, B. Emile, C. Rosenberger, et H. Laurent. *Unsupervised performance evaluation of image segmentation*. In *EURASIP Journal on Applied Signal Processing and Special Issue on Performance Evaluation in Image Processing*, 2006.

[Chabrier *et al.* 2006b] S. Chabrier, C. Rosenberger, H. Laurent, et Y.J. Zhang. *Adaptive evaluation of image segmentation results*. In *International Conference on Pattern Recognition (ICPR)*, 2006.

[Chapelle *et al.* 1999] O. Chapelle, P. Haffner, et V. Vapnik. *SVMs for Histogram-Based Image Classification*. IEEE Transaction on Neural Networks, 10(5) : 1055–1064, 1999.

[Chen *et al.* 2003] Tzong-Jer Chen, Keh-Shih Chuang, Jay Wu, Sharon C Chan, Ing-Ming Hwang, et Meei-Ling Jan. *A novel image quality index using Moran I statistics*. Physics in Medicine and Biology, 48 : 131–137, 2003.

[Cocquerez et Devars 1985] J-P. Cocquerez et J. Devars. *Détection de Contours dans les Images Aériennes : Nouveaux Opérateurs*. Traitement du signal, 2(1) : 45–65, 1985.

[Cocquerez et Philipp 1995] J.-P. Cocquerez et S. Philipp. *Analyse d'Images : filtrage et segmentation*. Masson, 1995.

[Comaniciu et Meer 2002] Dorin Comaniciu et Peter Meer. *Mean Shift : A Robust Approach Toward Feature Space Analysis*. IEEE Transactions on Pattern analysis and Machine Intelligence, 24 (5) : 603–619, 2002.

[Coquin *et al.* 1997] D. Coquin, P. Bolon, et Y. Chehadeh. *Evaluation quantitative d'images filtrées*. In *GRETSI*, volume 2, pages 1351–1354, 1997.

[Coquin et Bolon 2001] D. Coquin et Ph. Bolon. *Application of Baddeley's distance to dissimilarity measurement between gray scale images*. Pattern Recognition Letters, 22(14) : 1483–1502, 2001.

[Correia et Pereira 2000] P. Correia et F. Pereira. *Objective Evaluation of Relative Segmentation Quality*. International Conference on Image Processing, pages 308–311, September 2000.

[Courreges *et al.* 2001] F. Courreges, N. Smith, G. Poisson, P. Vieyres, A. Gourdon, M. Szpieg, et O. Merigeaux. *Real-time exhibition of a simulated space tele-echography using an ultra-light robot*. I-SAIRAS, juin 2001.

[Cryer et Shah 1999] J.E. Ruo Zhang Ping-Sing Tsai Cryer et M. Shah. *Shape-from-shading : a survey*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 21(8) : 690–706, 1999.

[Daly 1992] S. Daly. *The Visible Differences Predictor : an Algorithm for the Assessment of Image Fidelity*. Proceedings of the SPIE and Human Vision and Visual Processing and Digital Display III, 1666 : 2–15, 1992.

[Delgorge *et al.* 2001] C. Delgorge, P. Vieyres, G. Poisson, C. Rosenberger, et P. Arbeille. *Comparative survey of ultrasound images compression methods dedicated to a tele-echography robotic system*. 23rd Annual Conference IEEE/EMBS Engineering in Medicine and Biology Society, octobre 2001.

[Delgorge *et al.* 2002] C. Delgorge, C. Rosenberger, P. Vieyres, et G. Poisson. *JPEG 2000 and an adapted compression method for ultrasound images ? A comparative study*. In *International Conference on Science and Informatics*, volume 9, pages 536–539, juillet 2002.

[Delgorge *et al.* 2005a] C. Delgorge, F. Courreges, L. Al Bassit, C. Novales, C. Rosenberger, N. Smith-Guerin, C. Bru, R. Gilabert, M. Vannoni, G. Poisson, et P. Vieyres. *A tele-operated mobile ultrasound scanner using a light weight robot*. IEEE Transactions on Information Technology in BioMedicine and special issue mHealth, 9(1), 2005.

[Delgorge *et al.* 2005b] C. Delgorge, C. Rosenberger, G. Poisson, et P. Vieyres. *Evaluation of the quality of ultrasound image compression by fusion of criteria with a genetic algorithm*. In *International Conference on Advances in Pattern Recognition (ICAPR)*, volume 2, pages 464–472. Lecture Notes in Computer science and ISBN 0302-9743, août 2005.

[Delgorge *et al.* 2005c] C. Delgorge, C. Rosenberger, A. Rakotomamonjy, G. Poisson, et P. Vieyres. *Evaluation of the quality of ultrasound image compression by fusion of criteria with a support vector machine*. EUSIPCO, pages CD–ROM, septembre 2005.

[Delgorge *et al.* 2005d] C. Delgorge, P. Vieyres, G. Poisson, et C. Rosenberger. *Evaluation of the Quality of Ultrasound Image Compression for a Robotic Tele-Echographic System*. In *International Conference on Computers in Cardiology (CINC)*, 2005.

[Everingham *et al.* 2006] M. Everingham, A. Zisserman, C. Williams, L. Van Gool, M. Allan, C. Bishop, O. Chapelle, N. Dalal, T. Deselaers, G. Dorko, S. Duffner, J. Eichhorn, J. Farquhar, M. Fritz, C. Garcia, T. Griffiths, F. Jurie, D. Keysers, M. Koskela, J. Laaksonen, D. Larlus, B. Leibe, H. Meng, H. Ney, B. Schiele, C. Schmid, E. Seemann, J. Shawe-Taylor, A. Storkey, S. Szedmak, B. Triggs, I. Ulusoy, V. Viitaniemi, et J. Zhang. *The 2005 PASCAL Visual Object Classes Challenge*. In Selected Proceedings of the First PASCAL Challenges Workshop and LNAI, Springer-Verlag, 2006.

[Feng et Fulton 2005] D.D. He Chen Feng et R.S. Weidong Cai Fulton. *Performance evaluation of functional medical imaging compression via optimal sampling schedule designs and cluster analysis*. IEEE Transactions on Biomedical Engineering, 52 : 943 – 945, 2005.

[Forstner 1996] W. Forstner. *10 Pros and Cons against performance characterization of vision algorithms*. In *ECCV Workshop on Performance characterisitics of vision algorithms*, 1996.

[Gourdon *et al.* 1999] A. Gourdon, Ph. Poignet, G. Poisson, P. Vieyres, et P. Marché. *A new robotic mechanism for medical application*. In *IEEE International Conference on Advanced Intelligence Mecatronics, Atlanta, USA, Octobre 1999.*, 1999.

[Hall et Hall 1977] C.F. Hall et F. Hall. *A Nonlinear Model for the Spatial Characteristics of the Human Visual System*. IEEE Transaction on Systems and Man and Cybernetics, SMC-7(3) : 161–170, 1977.

[Han et Kim 2002] J. H. Han et T. Y. Kim. *Ambiguity distance : an edge evaluation measure using fuzziness of edges*. Fuzzy Sets and Systems, 126 : 311–324, 2002.

[Hao *et al.* 2003] P. Hao, Q. Shi, et Y. Chen. *Co-histogram and its application in remote sensing image compression evaluation*. In *IEEE International Conference on Image Processing*, volume 2, pages 37–40, 2003.

[Haralick 1986] R.M. Haralick. *Computer vision theory : The lack thereof*. Computer Vision Graphics and Image Processing, 36 : 272–286, 1986.

[Haralick et Shapiro 1985] R. H. Haralick et L. G. Shapiro. *Image Segmentation Techniques*. Computer Vision Graphics Image Processing, 29 : 100–132, 1985.

[Havstad *et al.* 2004] A. Havstad, A. Bouzerdoum, et A. Beghdadi. *Image Quality Assessment using a Neural Network Approach*. In *Proceedings of Symposium on Signal Processing and Information Technology (ISSPIT)*, 2004.

[Holland 1975] J.H. Holland. *Adaption in Natural and Artificial Systems : an introductory analysis with applications to biology and control and artificial intelligence*. The MIT press / A bradford Book, Cambridge MA, 1975.

[Hoover *et al.* 1996] Adam Hoover, Gillian Jean-Baptiste, Xiaoyi Jiang, Patrick J. Flynn, Horst Bunke, Dmitry B. Goldgof, Kevin Bowyer, David W. Eggert, Andrew Fitzgibbon, et Robert B. Fisher. *An Experimental Comparison of Rande Image Segmentation Algorithms*. IEEE Transaction on Pattern Analysis and Machine Intelligence, 18(7) : 673–689, july 1996.

[Houck *et al.* 1995a] C. Houck, J. Jones, et M. Kay. *A comparison of genetic algorithms and random restart and and two-opt switching for solving large location-allocation problems*. Computers & Operations Research forthcoming in special issue on evolution computation, 1995.

[Houck *et al.* 1995b] C. R. Houck, J. A. Joines, et M. G. Kay. *A genetic algorithm for function optimisation : a ©matlab implementation.* Technical report, NCSU-IE, sept 1995.

[Huang et Dom 1995] Q. Huang et B. Dom. *Quantitative Methods of Evaluating Image Segmentation*. International Conference on Image Processing (ICIP'95) Washington DC, USA, 3 : 53–56, 1995.

[ITU-R-BT.500-11 2002] ITU-R-BT.500-11. *Methodology for the subjective assessment of the quality of television pictures*. Recommendation on Broadcasting Service, 2002.

[Kara-Falah 1995] R. Kara-Falah. *Segmentation d'images : coopération, fusion, évaluation*. Thèse de Doctorat, Université de Savoie, juin 1995.

[Kermad *et al.* 2000] C. Kermad, B. Vozel, et K. Chehdi. *Hyperspectral image analysis and dimensionality : a scalar scheme through multi-thresholding technique*. Proc of the Eos/Spie Symposium on Remote sensing, 31(4170), 2000.

[Kitchen et Rosenfeld 1984] L. Kitchen et A. Rosenfeld. *Scene analysis using region-based constraint filtering*. Pattern Recognition, 17(2) : 189–203, 1984.

[Klein 1993] S.A. Klein. *Image quality and image compression : a psychophysicist's viewpoint*. Digital and human vision, pages 73–88, 1993.

[Laurent *et al.* 2003] H. Laurent, S. Chabrier, B. Emile, et C. Rosenberger. *Etude comparative de critères d'évaluation de la segmentation*. GRETSI, 3 : 150–153, 2003.

[Lee et Lewicki 2002] Te-Won Lee et Michael S. Lewicki. *Unsupervised image classification, segmentation and enhancement usind ICA mixture models*. IEEE Transactions on Image Processing, 11 (3) : 270–279, 2002.

[Levine et Nazif 85] M. D. Levine et Ahmed M. Nazif. *Dynamic Measurement of Computer Generated Image Segmentations*. IEEE Transaction on Pattern Analysis and Machine intelligence, PAMI-7(2) : 155–164, March 85.

[Li *et al.* 2002] J. Li, G. Chen, et Z. Chi. *A fuzzy image metric with application to fractal coding*. IEEE Transactions on Image Processing, 11 : 636–643, 2002.

[Limb 1979] J.O. Limb. *Distorsion Criteria of the Human Viewer*. IEEE Transaction on Systems and Man and Cybernetics, SMC-9(12) : 778–793, 1979.

[Linfoot 1958] E.H. Linfoot. *Quality Evaluation of Optical Systems*. Optica Acta, 5(1-2) : 1–13, 1958.

[Liu et Yang 1994] J. Liu et Y.-H. Yang. *Multiresolution Color Image Segmentation*. IEEE Trans on PAMI, 16(7) : 689–700, 1994.

[López *et al.* 1998] V. Barranco López, R. Román Roldán, J.F. Gómez Lopera, et J. Martínez Aroza. *A measure of quality for evaluating methods of segmentation and edge detection*. International Computer Symposium, Workshop on Image Processing and Character Recognition ICS'98, pages 221–226, 1998.

[Lorigo et Govindaraju 2006] L.M. Lorigo et V. Govindaraju. *Offline Arabic Handwriting Recognition : A Survey*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 28(5) : 712–724, 2006.

[Mannos et Sakrison 1974] J.L. Mannos et D.J. Sakrison. *The Effects of Visual Fidelity Criterion on the Encoding of Images*. IEEE Transactions on Information Theory, IT-20(4) : 525–536, 1974.

[Martin *et al.* 2001] D. Martin, C. Fowlkes, D. Tal, et J. Malik. *A Database of Human Segmented Natural Images and its Application to Evaluating Segmentation Algorithms and Measuring Ecological Statistics*. In *8th International Conference on Computer Vision*, pages 416–423, July 2001.

[Marzouki *et al.* 1997] A. Marzouki, Y. Delignon, et W. Pieczynski. *Estimation of generalized mixtures and its application in image segmentation*. IEEE Transactions on Image Processing, 6(10) : 1364–1375, octobre 1997.

[Michel *et al.* 1994] O. Michel, R. G. Baraniuk, et P. Flandrin. *Time frequency based distance and divergence measures*. In *IEEE International Symposium on TF and TS analysis*, pages 64–67, 1994.

[Miyahara *et al.* 1996] M. Miyahara, K. Kotani, et V. R. Algazi. *Objective Picture Quality Scale (PQS) for Image Coding*. IEEE Transactions on Communications, 46 : 1215–1226, 1996.

[Miyaji *et al.* 2000] Satoshi Miyaji, Takahiro Hamada, et Shuichi Matsumoto. *Development of a Digital Compressed Picture Quality Assessment System Considering Human Visual Perception*. Systems and Computers in Japan, 31(5) : 97–109, 2000.

[Naegele-Jackson *et al.* 2002] S. Naegele-Jackson, P. Holleczek, T. Rabenstein, J. Maiss, et M. Sackmann. *Influence of compression and network impairments on the picture quality of video transmissions in tele-medicine*. In *Proceedings of the 35th Annual Hawaii International Conference on System Sciences*, pages 2060 – 2068, 2002.

[Odet *et al.* 2002] C. Odet, B. Belaroussi, et H. Benoit-Cattin. *Scalable Discrepancy Measures for Segmentation Evaluation*. ICIP, 1 : 785–788, septembre 2002.

[Olivès 1998] Jean-Luc Olivès. *Optimisation globale d'un système imageur à l'aide de critères de qualité visuelle*. Thèse de Doctorat, Ecole Nationale Supérieure de l'Aéronautique et de l'Espace ENSAE, septembre 1998.

[Osuna *et al.* 1997] E. Osuna, R. Freund, et F. Girosi. *Training support vector machines : An application to face detection*. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 1997.

[Pavlovic *et al.* 1997] V.I. Pavlovic, R. Sharma, et T.S. Huang. *Visual interpretation of hand gestures for human-computer interaction : a review*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 19(7) : 677–695, 1997.

[Perlmutter *et al.* 1998] Sharon M. Perlmutter, Pamela C. Cosman, Chien-Wen Tseng, Richard A. Olshen, Robert M. Gray, King C. P. Li, et Colleen J. Bergin. *Medical Image Compression and Vector Quantization*. Statistical Science, 13(1) : 30–53, 1998.

[Philipp-Foliguet et Guigues 2005] S. Philipp-Foliguet et L. Guigues. *Critères multi-échelles d'évaluation de la segmentation*. GRETSI, 2005.

[Pommert et Höhne 2002] Andreas Pommert et Karl Heinz Höhne. *Evaluation of Image Quality in Medical Volume Visualization : The State of Art*. Lecture Notes in Computer Science, 2489 : 598–605, 2002.

[Pratt *et al.* 1978] W. Pratt, O. D. Faugeras, et A. Gagalowicz. *Visual discrimination of stochastic texture fields*. IEEE Transactions on Systems, Man, and Cybernetics, 8(11) : 796–804, 1978.

[Price 1986] K.E. Price. *Anything i can do and i can do better, no you can't*. Computer Vision Graphics and Image Processing, 36 : 387–391, 1986.

[Rakotomamonjy 2003] A. Rakotomamonjy. *Variable selection using SVM-based criteria*. Journal of Machine Learning Research and Special Issue on Variable Selection, 3 : 1357–1370, 2003.

[Roman-Roldan *et al.* 2001] R. Roman-Roldan, J. F. Gomez-Lopera, C. Atae-Allah, J. Martinez-Aroza, et P. L. Luque-Escamilla. *A measure of quality for evaluating methods of segmentation and edge detection*. Pattern Recognition, 34 : 969–980, 2001.

[Rosenberger 1999] C. Rosenberger. *Mise en Oeuvre d'un Système Adaptatif de Segmentation d'Images*. Thèse de Doctorat, Université de Rennes 1, decembre 1999.

[Rosenberger et Chehdi 2002] C. Rosenberger et K. Chehdi. *Supervised Genetic Image Segmentation*. In *IEEE International Conference on System, Man and Cybernetics*, 2002.

[Rosenberger et Chehdi 2003a] C. Rosenberger et K. Chehdi. *Towards a complete adaptive analysis of an image*. International Journal of Electronic Imaging, 12(2) : 292–298, 2003.

[Rosenberger et Chehdi 2003b] C. Rosenberger et K. Chehdi. *Unsupervised segmentation of multi-spectral images*. In *International Conference on Advanced Concepts for Intelligent Vision Systems*, 2003.

[Rosenfeld et Kak 1976] Azriel Rosenfeld et Avinash C. Kak. *Digital Picture Processing*, volume I. Academic Press, New-York, 1976.

[Schettini *et al.* 2002] Raimondo Schettini, Gianluigi Ciocca, et Silvia Zuffi. *Color Imaging Science : Exploiting Digital Media*, chapter A survey of methods for colour image indexing and retrieval in image databases, pages 183–211. R. Luo, L. MacDonald eds., J. Wiley, 2002.

[Shade 1956] O.H. Shade. *Optical and Photoelectric Analog of the Eye*. Journal of the Optical Society of America, 46(9) : 721–739, 1956.

[Silvestre *et al.* 2004] J. Silvestre, V. Sempere, et T. Albero. *Industrial video sequences for network performance evaluation*. IEEE International Workshop on Factory Communication Systems, pages 343 – 347, 2004.

[Simard *et al.* 1993] P. Simard, Y. LeCun, et J. Denker. *Efficient pattern recognition using a new transformation distance*. In Morgan Kaufman, editor, *Advances in Neural Information Processing Systems*, volume 5, pages 50–58, 1993.

[Singh et Amin 1999] S. Singh et A. Amin. *Neural Network Recognition of Hand-printed Characters*. Neural Computing and Applications, 8(1) : 67–76, 1999.

[Sundersingh 2000] Bijoy J. Sundersingh. *Qualitative Evaluation of Data Compression in Real-Time Ultrasound Imaging*. Thèse de Doctorat, University of Tennessee, Memphis, juin 2000.

[Tamtaoui *et al.* 1999] Ahmed Tamtaoui, Marc Ribagnac, et Driss Aboutajdine. *Elaboration de critères de mesure de qualité objective pour la qualité subjective*. In *GRETSI*, pages 845–848, septembre 1999.

[Thacker *et al.* 2005] N.A. Thacker, A.F. Clark, J. Barron, R. Beverige, C. Clark, P. Courtney, W.R. Crum, et V. Ramesh. *Performance characterization in Computer Vision*. , University of Manchester, 2005.

[Tsai *et al.* 2005] A. Tsai, A. Yezzi, W. Wells, C. Tempany, D. Tucker, A. Fan, W.E. Grimson, et A. Willsky. *A shape-based approach to the segmentation of medical imagery using level sets*. IEEE Transactions on Medical Imaging, 22 : 137–154, 2005.

[Turaga *et al.* 2004] Deepak S. Turaga, Yingwei Chen, et Jorge Caviedes. *No reference PSNR estimation for compressed pictures*. Signal Processing : Image Communication, 19 : 173–184, 2004.

[Vapnik 1998] V. Vapnik. *Statistical Learning Theory*. Wiley, 1998.

[Vinet 1991] L. Vinet. *Segmentation et mise en correspondance de régions de paires d'images stéréoscopiques*. Thèse de Doctorat, Université de Paris IX Dauphine, Juillet 1991.

[Wang *et al.* 2004] Zhou Wang, Alan Conrad Bovik, Hamid Rahim Sheikh, et Eero P. Simoncelli. *Image Quality Assessment : From Error Visibility to Structural Similarity*. IEEE Transactions on Image Processing, 13(4) : 600–612, avril 2004.

[Weatherburn *et al.* 2003] Gwyneth C. Weatherburn, Deborah Ridout, Nicola H. Strickland, Peter Robins, Christine M. Glastonbury, Walter Curati, Chris Harvey, et Clair Shadbolt. *A comparison of conventional film and CR hard copy and PACS soft copy images of the chest : analyses of ROC curves and inter-observer agreement*. European Journal of Radiology, 47 : 206–214, 2003.

[Weszka et Rosenfeld 1978] J. S. Weszka et A. Rosenfeld. *Threshold evaluation techniques*. IEEE Transactions on Systems, man and cybernetics, 8 : 622–629, 1978.

[Wilson *et al.* 1997] D.L. Wilson, A.J. Baddeley, et R.A. Owens. *A new metric for grayscale image comparison*. Intenration journal on Computer Vision, 24 : 5–17, 1997.

[Winkler 2005] Stefan Winkler. *Digital video quality : Visions models and metrics*. John Wiley & sons, 2005.

[Yang 2005] S.-B. Yang. *Smooth side-match weighted vector quantizer with variable block size for image coding*. IEE Proceedings Vision and Image and Signal Processing, 152 : 763–770, 2005.

[Yasnoff *et al.* 1977] W. A. Yasnoff, J. K. Mui, et J. W. Bacus. *Error measures for scene segmentation*. Pattern Recognition, 9 : 217–231, 1977.

[Yoshimura et Oe 1999] Motohide Yoshimura et Shunichiro Oe. *Evolutionary Segmentation of Texture Image Using Genetic Algorithms Toward Automatic decision of Optimum Number of Segmentation Areas*. Pattern recognition, 32 : 2041–2054, 1999.

[Zachariadis *et al.* 2002] K.E. Zachariadis, N.V. Boulgouris, N. Thomos, G.A. Triantafyllidis, et M.G. Strintzis. *Wavelet-based Communication of Medical Image Sequences*. 4th International Workshop on Enterprise Networking and Computing in Health Care, juin 2002.

[Zamperoni et Starovoitov 1996] P. Zamperoni et V. Starovoitov. *On measures of dissimilarity between arbitrary gray-scale images*. International Journal of Shape Modeling, 2 : 189–213, 1996.

[Zeboudj 1988] R. Zeboudj. *Filtrage, Seuillage Automatique, Contraste et Contours : du Pré-Traitement à l'Analyse d'image*. Thèse de Doctorat, Université de Saint Etienne, 1988.

[Zhang 1997] Y.J. Zhang. *Evaluation and comparison of different segmentation algorithms*. Pattern Recognition Letters, 18 : 963–974, 1997.

[Zhong et Ning 2005] J. Zhong et R. Ning. *Image denoising based on wavelets and multifractals for singularity detection*. IEEE Transactions on Image Processing, 14 : 1435–1447, 2005.