INSTITUT NATIONAL DES SCIENCES APPLIQUEES DE LYON

Année 2006 $N°$

THÈSE

pour obtenir le grade de
DOCTEUR
en
INFORMATIQUE

préparée au sein de l'école doctorale
INFORMATIQUE ET INFORMATION POUR LA SOCIÉTÉ

présentée et soutenue publiquement par

Nicolas ZLATOFF

le 12 JUILLET 2006

Indexation d'images 2D Vers une reconnaissance d'objets multi-critères

préparée au sein du laboratoire LIRIS
sous la direction de
Atilla Baskurt et Bruno Tellez

COMPOSITION DU JURY

Mme	Catherine Berrut	Rapporteur	Professeur
Mme	Christine Fernandez-Maloigne	Rapporteur	Professeur
M.	Jacques Blanc-Talon	Examinateur	DGA/D4S/MRIS
M.	Pascal Guitton	Examinateur	Professeur
M.	Atilla Baskurt	Directeur de thèse	Professeur
M.	Bruno Tellez	Co-directeur de thèse	Maitre de Conférences
M.	Noel Richard	Membre invité	Maitre de Conférences

Laboratoire d'InfoRmatique en Images et Systèmes d'information, UMR 5205 CNRS

Résumé

D'importants volumes d'images numériques, conduisent aujourd'hui à une forte demande d'outils permettant d'indexer puis de rechercher une image. Indexer une image consiste à en extraire une signature. Rechercher une image dans une base consiste alors à comparer plusieurs signatures entre elles. Une indexation est dite basée sur le contenu lorsqu'elle utilise les données de bas niveau (couleur, texture) de l'image pour construire la signature. De tels système sont faces à une limitation fondamentale : ils permettent aux utilisateurs de rechercher des images d'après leurs caractéristiques de bas niveaux (matière) alors ces derniers préfèreraient une recherche plus sémantique, relative à ce que l'image décrit (les objets présents, par exemple).

Dans cette thèse, nous proposons un système d'indexation qui permet de réduire le fossé entre les données de bas niveau et la sémantique. Tout d'abord, l'utilisateur formule, lors de la requête, un modèle (prototype) de l'objet recherché. Lors de la comparaison, entre ce modèle et les images de la base, plusieurs critères sont utilisés, comme la forme mais aussi l'organisation spatiale de différentes zones d'intérêt.

Une étape cruciale consiste justement à extraire de telles zones d'intérêt. Les approches de segmentation sont souvent entachées d'erreur, notamment à cause de variation d'éclairage dans la scène. Nous proposons donc de ne pas décrire une image par une segmentation unique mais plutôt par une hiérarchie de segmentations. Celle-ci représente l'image à différents niveaux de détails et se construit à partir de regroupements successifs de régions (groupements perceptuels), basés à la fois sur des critères de bas niveaux mais aussi géométriques.

Durant la comparaison entre un modèle et une image, nous considérons les correspondances entre chacune des parties au lieu d'utiliser seulement le modèle dans sa globalité. Plus précisément, la correspondance prend en compte les formes des parties, à travers les descripteurs ART (Angular Radial Transform) et CSS (Curvature Scale Space). En outre, l'organisation spatiale des parties entre elles est également prise en compte. Toutes ces caractéristiques sont combinées entre elles, par la théorie de l'évidence de Shafer afin d'en déduire une mesure unique de similarité.

Mots clés : indexation, image, segmentation, théorie de l'évidence, similarité.

Abstract

Huge volume of numeric images has recently led to strong needs for indexing and retrieval tools. Indexing an image consists in extracting a signature from it. Then, retrieving an image from an image database implies to compare several signatures together. We call content-based image retrieval systems those which build a signature from image low-level signal features such as color or texture. Such systems face a crucial limitation today. As a matter of fact, they allow to retrieve an image based on signal point of view, while users usually seek a more semantic-based search, related to what the image depicts (objects for instance).

In this thesis, we have proposed an indexing system which may allow to bridge the gap between low-level features and semantic. First, the user has to formulate a kind of model (prototype) for the object sought. Then, while comparing this model whith each image from the database, several features are considered, such as shape but also structural relationships between some regions of interest.

The extraction of those regions remains an open and challenging problem. Segmentation approaches are often error-prone, because of artifacts from tight variations in illumination of the scene. That is why we do not describe an image with one unique segmentation, but rather with a hierarchy of segmentations. This represents the image at several levels of detail. It is build by iterative perceptual groupings on regions, considering both low-level and geometric features.

When comparing a model with an image, we use one-to-one matching between model parts and regions from image, instead of considering the model in its whole. More precisally, comparison is based on shape similarity (through Angular Radial Transform and Curvature Scale Space) and on structural relationships among parts of object. All these features are then combined together, using Dempster-Shafer theory of belief, in order to derive one single similarity measure.

Keywords : indexing, image, segmentation, Dempster-Shafer theory, similarity.

Table des matières

Table des figures

Introduction

L'avènement de l'ère du "tout numérique" ces dernières décennies a conduit à une augmentation considérable du nombre d'images numériques disponibles en ligne. Un faisceau de causes permet d'expliquer cette situation. Peuvent notamment être identifiées, la très large utilisation des appareils photos numériques, l'augmentation des capacités de stockage des ordinateurs personnels et des serveurs, ou encore la multiplication des connexions haut-débit qui permettent d'échanger beaucoup plus facilement des données multimédia.

Quoi qu'il en soit, ce volume considérable d'images numériques est aujourd'hui une réalité. A titre d'illustration, Lyman et Varian (2003) annoncaient il y a quelques années déjà qu'il existait 4 milliards de sites internet, que ce nombre augmentait de 7,3 millions par jours, et qu'un site internet possédait en moyenne 14 photos. Ceci donne une idée du nombre démesuré d'images numériques disponibles aujourd'hui.

De ce contexte émerge logiquement une forte demande d'outils permettant de manipuler ces données. Ainsi, un utilisateur souhaitera classer puis retrouver ses images, ou encore parcourir des collections dans leur intégralité ou par bribes. Au coeur de cette manipulation de données, on trouve le processus dit *d'indexation*. Sans entrer pour l'instant dans le détail, on peut d'ores et déjà définir l'acte d'indexer un document, quel qu'il soit, comme le fait de le décrire, en vue de le retrouver plus tard.
Ainsi, indexer une image dans une base, consiste à stocker une information décrivant cette image. Une telle information est appelée index, et peut être de forme variée, comme nous le verrons par la suite. Par exemple, l'index le plus intuitif pour une image est sans doute celui constitué d'un mot-clé.

Une fois que toute une collection d'images a été indexée, se pose la question de la recherche d'information proprement dite : un utilisateur, qui recherche une ou des images, formule une *requête* au système d'indexation. Encore une fois, la forme de cette requête est variable. En particulier, elle est fortement conditionnée par la forme de l'index utilisé. Si l'on reprend l'exemple précédent où une image était indexée par un mot-clé, une requête peut consister en une série de mots-clés. Nous verrons que, dans le domaine des images, un autre type de requête est très répandu : celui de la requête par l'exemple. Dans ce cas, la requête est elle-même une image et l'utilisateur demande au système de retrouver les images jugées similaires à cet exemple.

Enfin, étant donnée une requête, le système doit chercher, sur la base des indexs dont il dispose, s'il existe un ou des documents jugés pertinents en réponse à la requête et les afficher à l'utilisateur.

La figure 1 résume les différentes étapes du processus : indexation, requête, traitement et réponse. Notons que la phase d'indexation proprement dite peut être réalisée hors-ligne alors que les autres impliquent un traitement en temps réel.

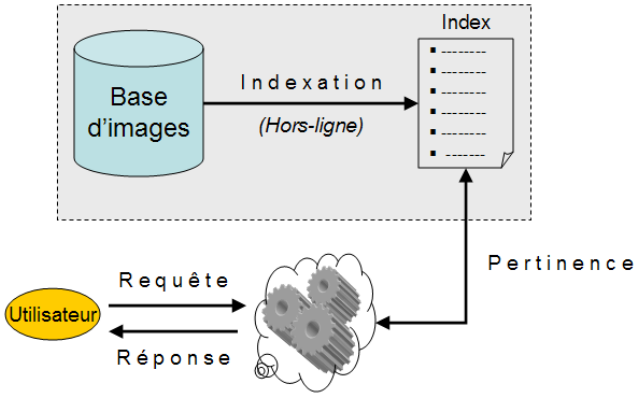


FIG. 1 – Principe général de l'indexation

Nous verrons dans l'état de l'art sur l'indexation qu'il existe deux grandes "familles" de méthodes pour l'indexation d'images. La première consiste à décrire l'image d'après ce qu'elle représente, avec des mots, ou plus généralement des labels sémantiques. La seconde méthode consiste à extraire de l'image des données relatives à son signal (couleur, texture, forme, etc...) et à utiliser ces dernières directement en tant qu'index. Issus d'une communauté de recherche clairement orientée vers la seconde méthode, nous nous engagerons très naturellement dans cette voie. Néanmoins, forts des interactions que nous avons pu nourrir avec différentes communautés, nous commenterons souvent nos résultats en regard des approches issues de la première famille. Ceci nous permettra de situer nos travaux dans une perspective élargie sur l'indexation, sans nous cloisonner à une seule communauté.
En outre, nous limiterons notre étude à des images fixes, à deux dimensions.

Nous le verrons dans le chapitre 1 de cette thèse, une étape cruciale de l'indexation consiste à partitionner l'image en différentes zones. Dans le cas idéal, ces zones correspondraient exactement aux différents objets qui y sont décrits. Il serait alors possible d'extraire différents descripteurs (couleur, forme, etc...) sur chacune des zones et de décrire ainsi une image comme une composition d'objets possédant chacun une liste de caractéristiques. Malheureusement, nous verrons que cette étape, appelée segmentation, est loin d'être triviale. C'est pourquoi nous lui consacrerons le chapitre 2, afin d'en étudier les mécanismes principaux ainsi que les limitations fondamentales. Habituellement, une segmentation utilise des caractéristiques locales comme la couleur ou la texture pour partitionner l'image en zones homogènes. Nous

proposerons dans ce document notre propre méthode de segmentation, qui utilise d'autres propriétés pour la partition. Ces dernières sont issues de travaux psychovisuels sur la perception humaine.

Néanmoins, dans un cadre très général, il est pratiquement impossible d'obtenir une segmentation robuste, c'est-à-dire qui permette de découper l'image en zones correspondant exactement aux objets qui y sont décrits. Nous étudierons alors comment il est possible d'utiliser une segmentation "imparfaite" en vue d'extraire différents descripteurs, notamment liés à la forme, afin d'indexer chaque partie d'image. En particulier nous insisterons sur l'utilité de manipuler une hiérarchie de segmentations (section 3.1.2 du chapitre 3).

A ce stade, il s'agira alors de savoir comment combiner différents descripteurs (couleur, texture, forme) afin d'évaluer une mesure de similarité unique et globale entre la requête soumise et chaque image indexée. C'est ainsi que la section 3.4 du chapitre 3 sera dédiée au mécanisme de requête proprement dit. Nous montrerons en particulier comment nous utilisons un formalisme mathématique, celui de la théorie de l'évidence, pour pouvoir combiner l'influence de différents descripteurs lors de la requête.

Enfin, un dernier chapitre nous permettra de conclure cette thèse et d'ouvrir différentes perspectives de travaux futurs.

1

Etat de l'art en indexation d'images

Sommaire

1.1 Introduction

1.1.1 A propos des usages

S'il est aujourd'hui largement reconnu que la question de l'accès aux bases d'images constitue une problématique importante de recherche, la question des modes opératoires est, elle, beaucoup moins claire. En effet, les usages souhaités par les utilisateurs des systèmes d'indexation sont peu ou mal définis. Sans entrer dans le détail, nous présenterons ici trois cas d'utilisation basiques, d'après Smeulders et al. (2000) et Santini (2001) : les recherches par cible, par catégorie et par association.

1.1.1.1 Recherche par cible

L'utilisateur a déjà une idée plus ou moins précise de ce qu'il cherche. Deux cas peuvent se présenter. L'image cherchée doit contenir un objet particulier ou alors doit correspondre à une *forme mentale* précise. Par ce terme, nous entendons une représentation mentale du type "un paysage avec une zone verte au tiers inférieur et un ciel bleu au-dessus".

Les système d'indexation basés contenu dont nous reparlerons plus loin se sont largement focalisés sur ce type de recherche, notamment en faisant un recours quasi-systématique au paradigme de requête par l'exemple : "étant donnée une image requête, je recherche une image similaire". On voit d'emblée la limitation de ce type d'approche : comment le système peut-il inférer ce que cherche l'utilisateur à partir de la requête ? Est-ce le même objet que celui présent dans l'image requête ? Est-ce une image présentant une même distribution spatiale de couleur ? Nous reviendrons d'ailleurs sur ce point en examinant la notion même de similarité dans la partie 1.4.2.

1.1.1.2 Recherche par catégorie

Il s'agit d'une version plus large de la recherche précédente. Cette fois, on cherche un ensemble d'images qui représentent différentes instances d'une classe plus ou moins abstraite. Par exemple, si dans la recherche précédente par cible un utilisateur recherchait une image représentant l'objet "fraise", la version par catégorie de cette requête serait de rechercher des images de fruits.

Outre les problèmes inhérents à une requête classique par l'exemple, le problème fondamental ici est de pouvoir disposer de descripteurs suffisament expressifs sur les images pour pouvoir rester insensibles aux variations d'aspects de plusieurs instances d'une même classe, tout en étant capable de séparer correctement deux instances de deux classes différentes.

1.1.1.3 Recherche par association

Cette fois, l'utilisateur n'a pas de but vraiment précis lorsqu'il lance sa recherche. L'exemple le plus couramment utilisé ici est celui d'un journaliste qui vient d'écrire un article théorique et qui cherche une image pour illustrer ce dernier. Beaucoup d'images peuvent convenir, chacune conférant une nouvelle portée à l'article.

Dans ce cas, la requête par l'exemple ne peut plus fonctionner. Au contraire, l'utilisateur a tout d'abord besoin de naviguer au sein de la base d'images pour en avoir un aperçu général. En outre, il est peu probable que l'utilisateur trouve ce qu'il cherche dès la première requête. Il faut donc prévoir des mécanismes d'interaction avec l'utilisateur, lui permettant d'affiner sa requête étape après étape.

1.1.2 A propos des approches d'indexation

Maintenant que nous avons rapidement passé en revue les usages d'un système d'indexation, penchons-nous sur le processus d'indexation proprement dit. On constate deux approches à l'indexation (Jorgensen, 2003). La première est centrée sur l'humain et décrit l'image d'après ce qu'elle représente. La seconde est issue du domaine du traitement du signal et décrit une image d'après ses caractéristiques comme la couleur, ou la texture. On parle dans ce dernier cas d'indexation basée *contenu*.

Les deux parties suivantes détaillent ces paradigmes.

1.1.2.1 Indexation sémantique : paradigme

L'indexation sémantique consiste à décrire une image d'après ce qu'elle représente : sa sémantique. L'objet image est ici considéré non pas en tant que tel, mais plutôt par rapport à ce qu'il décrit, son sens, son interprétation.

Plus concrètement, l'image est décrite par un utilisateur humain. Le processus d'indexation consiste alors à déposer une annotation sur l'image. Dans le cas le plus simple, il s'agit d'un texte libre, ou d'une liste de mots-clés, comme dans la figure 1.1.

L'exemple illustre bien les deux problèmes principaux auxquels se heurte un opérateur humain lors de cette opération. D'une part, il faut choisir quel mot précis doit être employé pour l'annotation : faut-il laisser l'annotation *tigre* ou faut-il préciser l'espèce ? D'autre part, il s'agit de décrire différents aspects de l'image, comme le sujet, le fond, l'action... et une simple liste de mots-clés s'avère ambigüe. En effet, pour chaque mot-clé, on ne sait pas *de quoi* on parle. Par exemple, le mot *rivière* dans la figure 1.1 renvoie-t-il au sujet principal ou au décor ?

tigre, rivière, herbe, chasse

FIG. 1.1 – Exemple d'annotation manuelle.

La notion de thésaurus

Concernant la première question de savoir quel mot choisir pour chaque annotation, une solution est d'utiliser un vocabulaire contraint, c'est-à-dire une liste de mots pré-sélectionnés. On parle dans ce cas de thésaurus (ROUSSEY, 2001). Un certain nombre de thésaurus sont déjà utilisés dans des systèmes d'indexation d'images. Citons par exemple ICONCLASS, de VAN DERWAAL (1985), Professeur d'histoire de l'art, et appartenant actuellement à l'académie royale des arts et des sciences des Pays-Bas (KNAW). ICONCLASS propose 14000 mots-clés destinés à décrire des images sans contexte a priori. Les mots-clés sont organisés entre eux selon les relations *partie de* et *sorte de* selon dix classes principales : art abstrait, religion-magie, être humain, société-civilisation, idées, histoire, bible, littérature, mythologie.

Le Art and Architecture Thesaurus (AAT, (PETERSON, 1994)) propose quant à lui 125000 termes dédiés à la description d'objets relevant de l'art, de l'architecture ou plus généralement de la culture, depuis l'antiquité jusqu'au présent. Les mots sont ici organisés dans des hiérarchies *sorte de* ou *instance de*, selon sept facettes principales : concepts abstraits, attributs physiques, styles et périodes, agents, activités, matériels et objets.

La notion d'attributs - valeur

Concernant maintenant la deuxième question, à savoir qu'un mot seul est ambigü puisqu'on ne sait pas ce qu'il décrit, une solution consiste à structurer davantage les annotations. Par exemple, on peut utiliser des couples attributs - valeur. Chaque attribut est défini de façon plus ou moins formelle de manière à préciser sa sémantique. Une famille d'attributs très connue est le Dublin Core (Dublin Core Metadata Initiative, 2003). Elle est constituée de 15 attributs qui peuvent s'appliquer à tout type de document : titre, créateur, format, description par exemple. Une extension a été définie par le *Visual Resource Association* (VRA) pour la description d'objets artistiques, avec des attributs comme style, période, technique, etc...

Les travaux du web sémantique relèvent de cette approche. Ainsi, le langage RDF (*Resource Description Framework*), défini par le W3C (2004) permet de décrire tout type de ressource du web (dont les images) par des couples attributs valeur. Ces derniers peuvent être libres, ou contraints par des méta-descriptions. Le W3C (2004) définit également le langage RDFS (*Resource Description Framework Schema*) pour formaliser ces méta-descriptions. L'idée au coeur du web sémantique est que d'une part, n'importe qui peut créer des méta-descriptions avec RDFS et que d'autre part, tout le monde peut utiliser le schéma créé par quelqu'un d'autre pour annoter ses propres documents. Ceci place l'interopérabilité des annotations au coeur du système. La figure 1.2 donne un exemple d'annotation d'image au format RDF, sans méta-description RDFS.

<rdf :sujet : tigre>
<rdf :décors : rivière>
<rdf :décors : herbe>
<rdf :action : chasse>

FIG. 1.2 – Exemple d'annotation manuelle au format RDF.

Dans cette optique, LAFON et BOS (2000) propose un système d'annotations basiques d'images en RDF sur le web, selon un schéma défini en RDFS. Toutefois, ce schéma est très orienté méta-données (créateur de l'image, date, format de fichier...) et ne traite que très partiellement ce que l'image représente.

Notons qu'il est possible d'aller encore plus loin dans la structuration de l'annotation, en utilisant par exemple des graphes, qui permettent de mettre en relation les différents attributs des annotations (PRIÉ ET AL., 1999). Ainsi, dans l'exemple de la figure 1.2, l'attribut *tigre* pourra être mis en relation avec l'attribut *chasse*, pour indiquer que l'image décrit un tigre qui chasse.

La principale limitation de ces approches concerne leur coût de mise en oeuvre. L'indexation d'une image peut prendre de quelques dizaine de secondes à quelques heures suivant le degré de détail et la complexité des index. C'est pourquoi un deuxième type d'indexation est apparu, afin de chercher à automatiser le processus.

1.1.2.2 indexation basée contenu : paradigme

L'image est cette fois vue non plus comme ce qu'elle décrit mais comme un signal, porté par l'ensemble des pixels. Les techniques classiques de traitement du signal permettent alors d'extraire localement de l'image un certain nombre de descripteurs numériques, comme la couleur ou la texture des pixels. Ces descripteurs

vont être à l'origine d'un ensemble de traitements, qui aboutiront à la création d'une signature décrivant au mieux le contenu de l'image. On peut distinguer ceux-ci selon la partition qu'ils effectuent sur l'image pour le calcul de la signature : regroupement des descripteurs sur toute l'image, ou en régions caractéristiques, par exemple. Ces techniques seront développées dans la partie suivante.

La limitation principale de l'indexation par le contenu est qu'elle ne permet qu'une recherche d'images par des critères dits bas-niveaux comme la couleur ou la texture. Forsyth et al. (1997) parlent de recherche orientée *matériaux* par opposition à une recherche d'images orientée *choses* : par exemple, "je recherche des images représentant des voitures". Or, c'est justement cette dernière que souhaite, en général, l'utilisateur.

Une évolution naturelle des systèmes basés contenu consiste donc à tenter de dériver une interprétation de l'image à partir des descripteurs bas-niveaux. On retrouve en fait des problématiques similaires à des approches "vision" en informatique. Bien qu'un certain nombre de bons résultats aient été obtenus ainsi, il convient de bien rappeler deux limitations fondamentales des approches vision à partir d'images 2D, correspondant aux fossés sensoriel et sémantique (voir la figure 1.3).

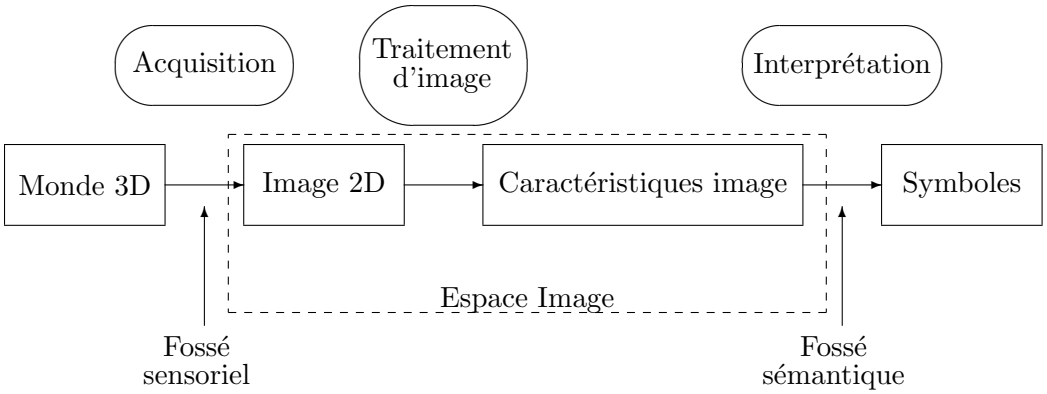


Fig. 1.3 – Les fossés sensoriel et sémantique.

Le fossé sensoriel rend compte de la nécessaire perte d'information résultant de la projection du monde réel 3D sur une image 2D. Outre des artefacts induits par le matériel de prise de vue (caméra, capteur, etc...), il s'agit plus significativement de la perte d'une dimension spatiale : la profondeur. Ceci fait de la vision un problème sous-contraint, ce qui limite donc fondamentalement ses résultats.

Le deuxième point, le fossé sémantique, rend compte de l'abscence de lien direct entre une série de descripteurs bas-niveaux extrait de l'image et une interprétation donnée. En effet, du fait par exemple du changement de couleur d'une surface en fonction de son orientation, un objet 3D du monde réel de couleur homogène peut très bien être représenté dans une image 2D par des zones de couleurs différentes. Ainsi, sa détection sur la seule base de critères bas-niveaux est fortement limitée.

1.1.2.3 Discussion sur la sémiotique

Il est communément admis que l'indexation sémantique manuelle, par un opérateur humain, est une opération fastidieuse et extrêmement coûteuse en temps. En outre, cette méthode pose la question de la subjectivité des indexs créés, puisqu'ils rendent compte du sens de l'image, capté par l'opérateur humain en cours d'indexation. C'est pourquoi la communauté de recherche "image", c'est-à-dire issue du traitement du signal, fustige souvent ces constats afin de justifier son approche basée contenu. En effet, cette dernière permet une automatisation des traitements ainsi que la création d'index "objectifs". Néanmoins, il faut bien voir que cette opposition est factice, car les deux approches ne considèrent en fait pas le même objet.

Une image peut être vue comme un *signe* c'est-à-dire comme un objet qui renvoie à une ou des interprétations. La sémiotique est la discipline qui étudie ce type d'objet. Ainsi que noté par Ferdinand de Saussure, un signe est composé de deux parties indissociables : un *signifiant* d'une part et un *signifié* d'autre part. Le signifiant renvoie à l'aspect matériel du signe. Dans le cas de l'image de la figure 1.4, il s'agit du signal couleur en deux dimensions porté par l'image. Le signifié est quant à lui une abstraction, généré par le signifiant. Ainsi, dans l'image de la figure 1.4, un des signifiés concerne le concept *tigre*, en tant que construction mentale.

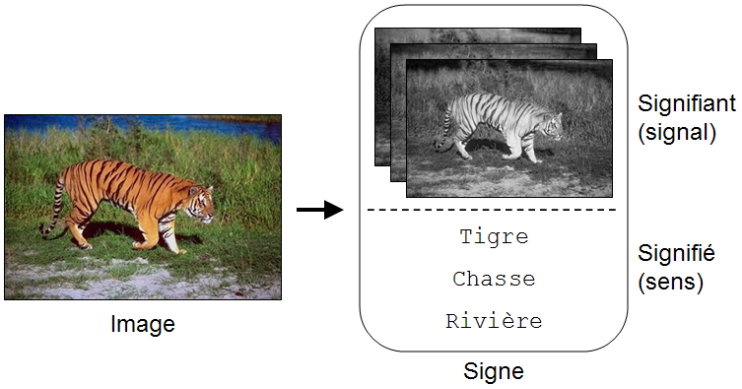


FIG. 1.4 – Une image vue comme un signe.

Il faut bien noter que signifié et signifiant sont indissociables dans un signe. En effet, un signifié seul n'est plus qu'un concept théorique sans lien avec notre réalité (il ne peut être nommé). De même, un signifiant seul n'est qu'un stimulus basique, qui ne renvoie à aucun concept connu.

Dans cette optique, il apparaît clairement que la méthode d'indexation manuelle se focalise sur la partie signifiée de l'image (sens), alors que la méthode basée contenu considère uniquement le signifiant : comment l'image représente son sens. Ainsi, les deux approches, loin de s'opposer, se complètent puisqu'elles décrivent chacune une partie d'un objet complexe, à deux facettes.

1.1.3 Positionnement et plan de l'état de l'art

Dans la suite de cette thèse, nous nous focaliserons sur l'indexation basée contenu, c'est-à-dire sur le signal porté par l'image. Néanmoins, nous garderons sans cesse à l'esprit qu'une image porte aussi un sens (signifié). Il s'agit donc de traiter le signal en tant que tel, de manière à obtenir des descriptions les plus sémantiques possibles, mais sans pour autant se substituer à l'indexation sémantique.

Nous allons maintenant dresser un état de l'art des approches d'indexation basée contenu. Au vue de la quantité de travaux sur la question, nous ne chercherons pas à être exhaustif. Au contraire, nous commenterons des approches caractéristiques, en regard du processus d'indexation.

Il est classiquement admis que le processus d'indexation basée contenu s'articule sur quatre étapes majeures. On distingue :

- Le traitement d'image (section 1.2) qui consiste à extraire en chaque point du signal porté par l'image, des données comme la couleur ou la texture.
- L'extraction de caractéristiques (section 1.3) consiste à regrouper les données précédentes sur certaines zones saillantes de l'image, afin d'extraire la signature de l'image.
- La requête (section 1.4) consiste ensuite à comparer une signature fournie par l'utilisateur (signature requête) à l'ensemble des signatures de la base, afin d'en déduire une liste d'images jugées pertinentes par rapport à ce que cherche l'utilisateur.
- Enfin, une dernière étape, optionnelle, concerne l'interaction du système avec l'utilisateur. Par exemple, si ce dernier n'est pas satisfait de la réponse du système, il peut avoir la possibilité d'affiner sa requête sous différentes modalités, afin d'avoir une réponse plus pertinente.

Nous allons maintenant détailler les trois premières étapes.

1.2 Description de contenu par traitement d'images

Il est important de noter que l'indexation basée contenu n'utilise pas l'image dans sa totalité lors du traitement d'une requête. Au contraire, elle repose sur des caractéristiques choisies, préalablement extraites. On peut considérer deux étapes (Smeulders et al., 2000) lors de l'extraction de ces caractéristiques sur une image $i(x, y)$. La première consiste à effectuer des opérations de traitement du signal, pour passer des données images à d'autres données spatiales, comme la couleur, la texture ou la géométrie locale. Elle peut être formalisée par :

$$f(x, y) = t \circ i(x, y) \tag{1.1}$$

avec t un opérateur de traitement d'images. La présente partie se concentre sur cette étape de traitement. Dans un deuxième temps, détaillé dans la partie suivante, les caractéristiques proprement dites de l'image seront extraites, à partir de ces nouvelles descriptions de couleur, géométrie et texture.

1.2.1 Couleur

La caractérisation de la couleur dans une image est une opération extrêmement complexe. En effet, cette donnée varie considérablement avec l'orientation des surfaces, le point de vue de la caméra et l'illumination (positions et longueur d'onde des sources lumineuses), par exemple. En outre, la perception de la couleur par l'être humain est un processus complexe et subjectif.

Il est possible de représenter la couleur dans différents espaces. Le plus répandu est sans aucun doute l'espace R, V, B (R, G, B en anglais) qui code la couleur d'un pixel sur un vecteur en trois dimensions : rouge, vert, bleu. La figure 1.5 illustre le codage d'une image avec ces trois composantes R, V, B, chacune des images en niveaux de gris représentant une composante couleur.

Image | Composante R | Composante V | Composante B

FIG. 1.5 – Exemple de description R, V, B.

Comme la plupart des images numériques code ainsi leur couleur, cet espace est fréquemment utilisé. En effet, aucune transformation n'est alors requise. En outre, cette modélisation paraît naturelle en ce qu'elle reflète le codage des couleurs par le système visuel humain.

Toutefois, cet espace possède quelques défauts importants. Tout d'abord, les différents canaux sont corrélés. En outre, la distance euclidienne entre deux points quelconques de cet espace ne rend pas forcément compte de la distance perceptuelle lue par un opérateur humain. C'est pourquoi d'autres espaces couleur ont été introduits. Citons parmi eux :

– H, S, V où chaque composante représente respectivement la teinte, la saturation et la luminance de la couleur. La composante H est invariante pour une orientation donnée de l'objet sous différentes illuminations, ce qui peut se révéler intéressant en indexation d'objets. La figure 1.6 illustre le codage d'une image avec ces trois composantes H, S, V, chacune des images en niveaux de gris représentant une composante.

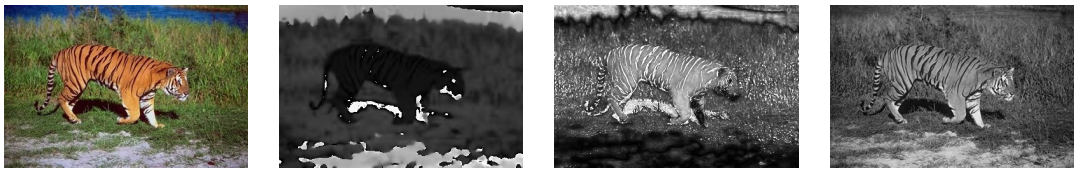
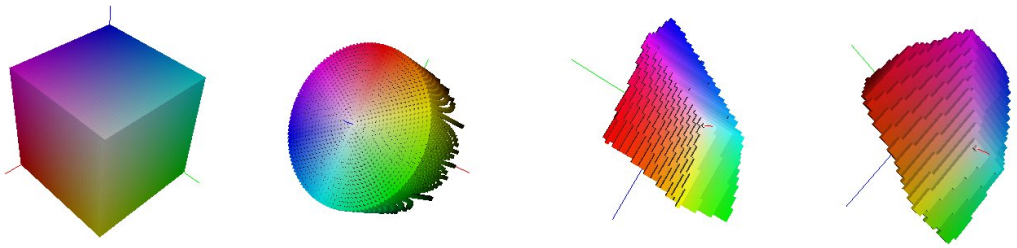


Image | Composante H | Composante S | Composante V

FIG. 1.6 – Exemple de description H, S, V.

– L, a, b ou L, u, v qui sont qualifiés d'homogènes : dans ces espaces, la distance euclidienne entre deux couleurs rend compte explicitement de la différence de perception des couleurs par l'être humain.

La figure 1.7 présente ces espaces R, G, B, H, S, V, L, a, b et L, u, v.



Espace R, G, B | Espace H, S, V | Espace L, a, b | Espace L, u, v

FIG. 1.7 – Exemple d'espaces couleurs.

1.2.2 Géométrie locale et texture

Par le terme *géométrie locale*, nous entendons les traitements qui décrivent les détails géométriques locaux d'une image. Les approches dites *différentielles* relèvent de la géométrie locale.

Il n'existe pas de définition précise et consensuelle du terme *texture* en vision par ordinateur. On parle fréquemment de répétitions de motifs similaires, sans pour autant que cette notion soit exhaustive. Nous avons donc choisi de regrouper dans un même paragraphe les descriptions de géométries locales et de textures.

1.2.2.1 Matrice de co-occurrence

Une approche classique de description de texture consiste à calculer une matrice de co-occurrence sur les niveaux de gris des pixels (HARALICK et SHAPIRO, 1992). Chaque élément P de position (i, j) dans cette matrice, rend compte de la probabilité qu'ont deux niveaux de gris i et j d'apparaître conjointement dans un

certain voisinage. Plusieurs statistiques peuvent alors être calculées sur la matrice de co-ocurrence. Citons, par exemple :

$$\text{l'énergie : } \sum_{i,j} (P(i,j))^2 \tag{1.2}$$

$$\text{l'entropie : } \sum_{i,j} P(i,j) \log P(i,j) \tag{1.3}$$

$$\text{le contraste : } \sum_{i,j} (i-j)^2 P(i,j) \tag{1.4}$$

$$\text{l'homogénéité : } \sum_{i,j} \frac{P(i,j)}{1+|i-j|} \tag{1.5}$$

Néanmoins, de telles approches sont limitées au traitement de zones de taille réduite, et ce, à cause de la taille de la matrice de co-occurrence.

1.2.2.2 Description fréquentielle : transformée de Fourier

D'autres approches proposent de passer dans un espace de fréquences spatiales, par exemple via une transformée de Fourier. Puisque les textures sont la répétition d'un motif, elles seront caractérisées par des pics dans la représentation fréquentielle de l'image (les texture fines dans les fréquences hautes et les textures grossières dans les basses fréquences). Etant donnée une image $i(x,y)$, la transformée de Fourier $F(f_x, f_y)$ s'exprime de la façon suivante :

$$F(f_x, f_y) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} i(x,y) e^{-2i\pi(f_x x + f_y y)} \, dx \, dy \tag{1.6}$$

f_x et f_y sont appelées les fréquences spatiales de l'image $i(x,y)$. On a, en outre, la relation inverse, dite de décomposition de la fonction image $i(x,y)$ sur une base de fonctions de Fourier :

$$i(x,y) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} F(f_x, f_y) e^{2i\pi(f_x x + f_y y)} \, df_x \, df_y \tag{1.7}$$

L'idée consiste donc à décrire l'image $i(x,y)$ comme étant un signal à deux variables spatiale x et y. Ce signal est décomposé comme une somme pondérée de fonctions sinusoïdales (équation 1.7). Le poids de chacune de ces fonctions dans la décomposition est rendue par le coefficient de Fourier $F(f_x, f_y)$. La figure 1.8 présente des exemples de textures, fines et grossières, ainsi que les modules $|F(f_x, f_y)|$ des transformées de Fourrier associées.

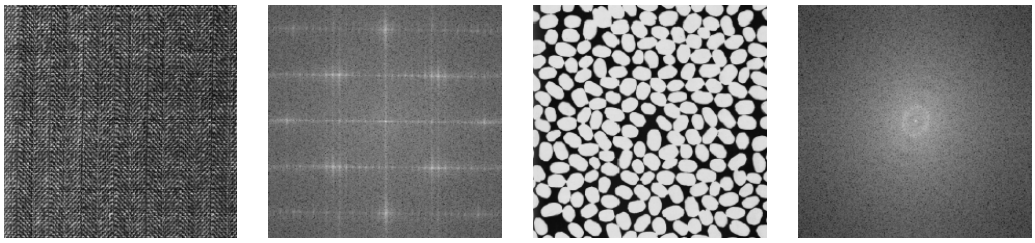


FIG. 1.8 – Exemple de texture et de module de la transformée de Fourier associée (horizontalement : f_x, verticalement : f_y)

Les descriptions de type Fourier décomposent l'image sur des fonctions sinusoïdales, qui ont la propriété de s'étendre indéfiniment dans l'espace. Une telle description est donc adaptée à des signaux périodiques ou quasi-périodiques, mais trouve sa limitation pour des signaux plus complexes, présentant de fortes discontinuités. Or, les images présentent justement ce type de caractéristiques. C'est pourquoi d'autres décompositions ont été utilisées, comme celle dite de Gabor.

1.2.2.3 Description espace - fréquence : méthode de Gabor

L'idée fondatrice a été de décomposer l'image sur des fonctions $G_{n,m}(x,y)$ limitées dans l'espace, afin d'en étudier indépendamment les fragments. On qualifie les $G_{n,m}(x,y)$ de fonctions analysantes. Dans le cas de la décomposition de Gabor, ces fonctions analysantes sont obtenues à partir d'une fonction mère $G(x,y)$, définie comme une fonction sinusoïdale orientée sur l'axe des x, modulée par une enveloppe gaussienne dans les directions x et y (voir figure 1.9(a)) :

$$G(x,y) = \frac{1}{2\pi\sigma_x\sigma_y} e^{-\frac{1}{2}\left(\frac{x^2}{\sigma_x^2}+\frac{y^2}{\sigma_y^2}\right)} e^{2i\pi f_x x} \tag{1.8}$$

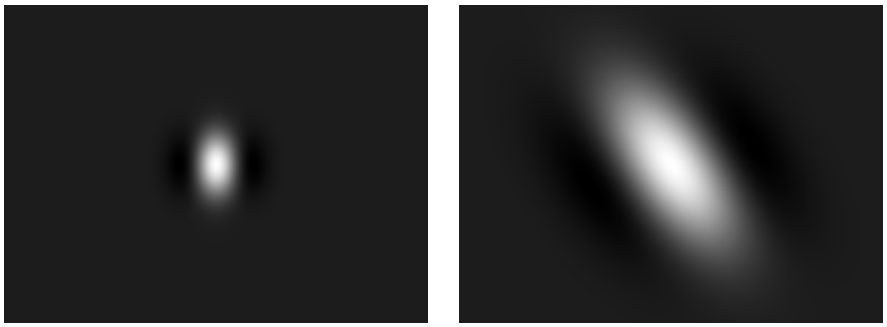
Les fonctions $G_{n,m}(x,y)$ sont ensuite obtenues par dilatation et rotation de $G(x,y)$:

$$G_{n,m}(x,y) = a^{-m} G(x',y') \tag{1.9}$$

$$x' = a^{-m}\left(x\cos\left(\frac{n\pi}{K}\right) + y\sin\left(\frac{n\pi}{K}\right)\right) \tag{1.10}$$

$$y' = a^{-m}\left(-x\sin\left(\frac{n\pi}{K}\right) + y\cos\left(\frac{n\pi}{K}\right)\right) \tag{1.11}$$

où a et K sont des paramètres à fixer. La figure 1.9(b) montre un exemple de dilation - rotation de la fonction mère $G_{n,m}(x,y)$.



(a) Fonction mère $G(x, y)$ (b) Fonction déformée $G_{n,m}(x, y)$

FIG. 1.9 – Exemple de filtre de Gabor

L'image i est alors analysée par cette série de fonctions, par convolution successives, conduisant à une série de coefficients $w_{n,m}(x, y)$:

$$w_{n,m}(x, y) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} G^*_{n,m}(x - a, y - a) i(a, b) \, da \, db \tag{1.12}$$

On parle de décomposition espace - fréquence, car les fonctions analysantes $(G_{n,m}(x, y))$ sont appliquées localement à l'image (espace) et à différentes fréquences du fait des dilatations et rotations. Cette décomposition a été utilisée par MANJUNATH et MA (1996) pour des indexations par les textures. TORRALBA et A.OLIVA (1999) utilisent également cette décomposition sur filtre de Gabor, afin de pouvoir classifier des images en familles sémantiques (paysage naturels, ville, intérieur, etc...), d'après les valeurs des coefficients $w_{n,m}(x, y)$.

Les filtres de Gabor sont relativement utilisés aujourd'hui, notamment du fait de leur pertinence en regard du système visuel humain. En effet, MARCELJA (1980) a montré que les cellules du cortex humain pouvaient être modélisées par des fonctions de Gabor à une dimension. Daugman a élargi ce modèle à deux dimensions.

Cependant, une limitation provient du choix non trivial des paramètres pour déformer la fonction mère analysante. En outre, la description obtenue via les coefficients $w_{n,m}(x, y)$ est redondante. Une classe plus générale de méthodes de description espace - fréquence est ainsi couramment utilisée en traitement d'images : les analyses multi-résolution par ondelettes.

1.2.2.4 Description espace - fréquence : ondelettes

L'idée fondatrice (MALLAT, 1999) est la même que pour les filtres de Gabor : décomposer l'image sur des fonctions $\Psi_{a,b}(x)$ limitées dans l'espace, afin d'en étudier les fragments indépendamment. Pour simplifier les notations, mais sans limitation aucune, plaçons-nous dans le cas d'un signal à une dimension. Une analyse par

ondelettes d'une fonction $i(x)$ consiste à projeter celle-ci sur une famille de fonctions $\Psi_{a,b}(x)$, appelées fonctions analysantes :

$$W(a,b) = \int_{-\infty}^{+\infty} i(x)\Psi_{a,b}(x)\,dx \tag{1.13}$$

Ces fonctions analysantes $\Psi_{a,b}(x)$ sont obtenues à partir de la déformation d'une fonction mère $\Psi(x)$, telle que :

$$\Psi_{a,b}(x) = \frac{1}{\sqrt{(a)}}\Psi\left(\frac{x-a}{b}\right) \tag{1.14}$$

Le paramètre a permet d'appliquer la fonction analysante en différent points de la fonction i (translation) et le paramètre b permet d'appliquer la fonction analysante à différentes échelles (dilatation de la fonction mère). La figure 1.10 présente deux déformées des ondelettes de Haar (en rouge). La fonction mère est affichée en bleue.

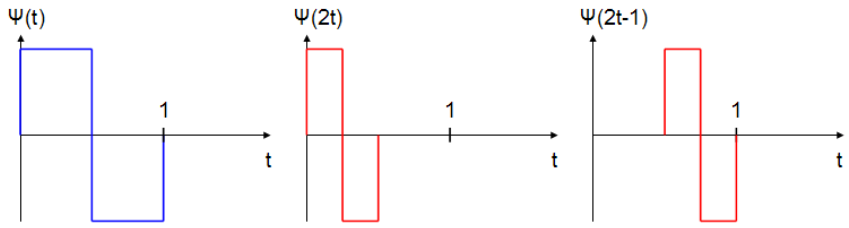
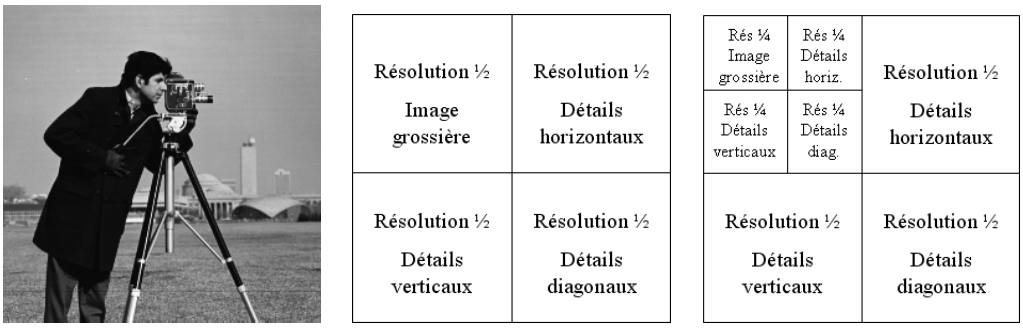


FIG. 1.10 – Ondelettes de Haar.

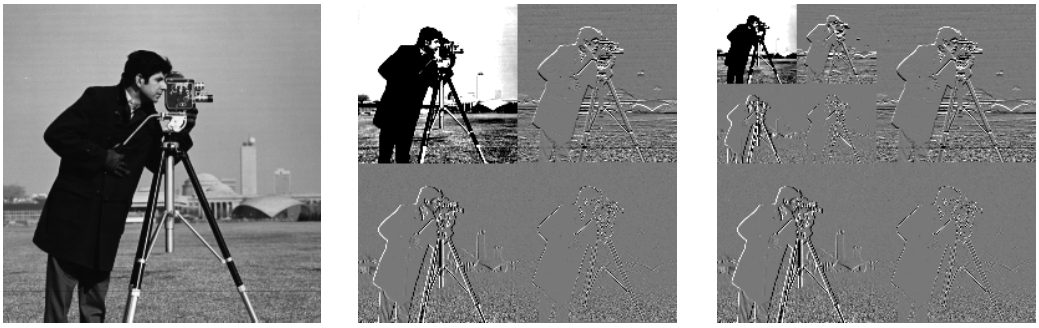
En pratique, la description par ondelettes d'une image, c'est-à-dire d'un signal à deux dimensions, traite séparemment chacune des composantes horizontale (x), verticale (y) et diagonale (x et y), sous-échantillonnées d'un facteur 2 par rapport à l'image originale. Il en résulte un schéma de décomposition à quatre cadrans, comme présenté dans la figure 1.11. Par exemple, lorsque la composante y est traitée, on obtient une image sous-échantillonnée d'un facteur 2, qui contient les détails verticaux (hautes fréquences, coin inférieur gauche). De même, le traitement de la composante x donne une image sous-échantillonnée qui contient les détails horizontaux. L'image originale peut alors être reformée, à partir des composantes basses-fréquences sur x et y (approximation grossière de l'image, coin supérieur gauche) auxquelles on ajoute les détails hautes fréquences des trois autres cadrans. L'opération ainsi réitérée est appelée analyse multi-résolution (AMR) par ondelettes.

La figure 1.12 montre un exemple de décomposition à deux niveaux d'une image, par les ondelettes de Haar.



(a) Image originale (b) AMR niveau 1 (c) AMR niveau 2

FIG. 1.11 – Principe de la décomposition AMR par ondelettes



(a) Image originale (b) AMR (Haar) niveau 1 (c) AMR (Haar) niveau 2

FIG. 1.12 – Décomposition AMR par ondelettes de Haar

1.2.3 Discussion sur le traitement d'images

Historiquement, les recherches se sont focalisées sur les aspects liés à la couleur, car celle-ci détient un pouvoir discriminant certain concernant les images. Les décompositions en temps-fréquence sont des travaux plus récents, les outils théoriques correspondants étant formalisés depuis peu.

On peut noter à ce stade que la distinction couleur - géométrie locale - texture est quelque peu artificielle. En effet, les descriptions ondelettes peuvent être étendues aux images couleurs. De même, les coefficients d'ondelettes véhiculent une information de texture. Comme noté par SMEULDERS ET AL. (2000), une description globale, qui regrouperait toutes ces notions pourrait s'avérer extrêmement riche en indexation. Pour l'instant, la plupart des systèmes décrivent une zone de l'image par un vecteur de caractéristiques renvoyant soit à la couleur, soit à la géométrie locale (CARSON ET AL., 2002; WANG ET AL., 2001).

Notons également que cette phase de traitement d'image, très bas-niveau, ne doit pas être négligée. En effet, elle peut jouer un rôle important dans la réduction de bruits ou autres artefacts de l'image : ombres, occlusion, bruits de capteurs de

caméra. Dès lors, les outils de traitements d'images permettent d'agir sur la réduction du fossé sensoriel.

1.3 Description de contenu par extraction de caractéristiques

Il s'agit maintenant de réduire l'espace de description, à partir des données $f(x, y)$ issues du traitement d'images. En effet, la signature de l'image doit réussir à synthétiser l'information pertinente de l'image en une série de caractéristiques. On peut classer celles-ci selon la partition qu'elles impliquent sur l'image pour leur calcul (Smeulders et al., 2000).

On peut ainsi distinguer :

- les partitions de l'image sur une grille fixe, indépendamment du contenu. Un cas limite consiste à n'utiliser aucune partition, c'est-à-dire à travailler sur toute l'image.
- les partitions par segmentation forte dans lesquelles l'image est découpée en régions qui correspondent extactement aux objets 3D du monde réel (image (d) de la figure 1.14). Notamment à cause des fossés sémantique et sensoriel, cette segmentation forte est impossible à obtenir sans l'assistance d'un utilisateur, dans un cas général, c'est-à-dire dans un univers non contraint. Toutefois, dans certains domaines (par exemple lorsque les conditions de prise de vue sont particulièrement strictes et constantes), cette partition est possible.
- les partitions par segmentation faible, qui constitue une solution dégradée de la segmentation forte. Cette fois, les régions extraites représentent une partition des objets réels. Un exemple de segmentation faible est visible sur l'image (c) de la figure 1.14. L'objet "perroquet" est en effet découpé en différentes zones à l'issue des traitements.

Chacune de ces partitions induit un type de caractéristiques privilégiées : par exemple, une partition de type grille fixe conduit à des descriptions par histogramme. Les parties suivantes détaillent les différents types de partitions.

1.3.1 Partition de type grille fixe et caractéristiques globales

Historiquement, ce sont les premières à avoir été utilisées. Elles conduisent à l'utilisation de caractéristiques globales, dont la plus connue est sans doute l'histogramme. Pour une image de taille $N * M$, en niveaux de gris, où chaque pixel est décrit par une composante discrète entière entre 0 et 255, l'histogramme se définit comme la fonction Hist qui à chaque niveau de gris n associe le nombre de pixels de

l'image possédant cette intensité : $card(n)$.

$$\begin{array}{rccc} \text{Hist} : & [0\,;255] & \longrightarrow & [0\,;N*M] \\ & n & & card(n) \end{array} \tag{1.15}$$

Réalisant que la couleur porte plus d'informations que le seul niveau de gris, Swain et Ballard (1991) ont montré qu'un histogramme de couleur peut se révéler un outil relativement adapté à l'indexation, si les structures recherchées possèdent une couleur unique. L'histogramme couleur d'une image est défini comme un histogramme à trois dimensions, où chaque dimension représente un histogramme sur une composante couleur. La figure 1.13 présente un exemple d'histogramme couleur sur une image, en considérant les trois composantes rouge-vert-bleu.

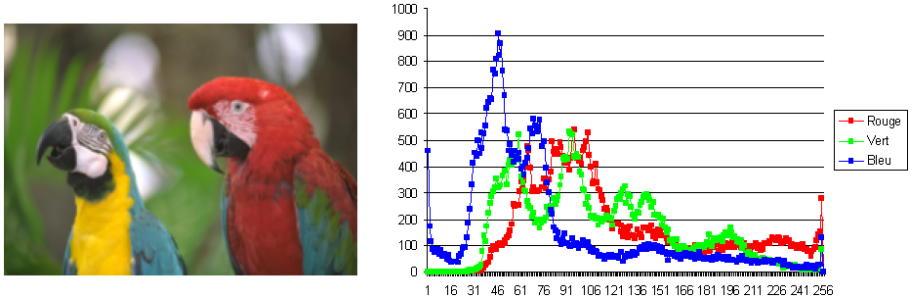


Fig. 1.13 – Exemple d'histogramme d'une image couleur.

Comme l'histogramme ne contient aucune information spatiale concernant la localisation d'un objet, une extension consiste à utiliser un corrélogramme (Huang et al., 1999). Celui-ci se définit comme un histogramme à trois dimensions, dans lequel les deux premières dimensions représentent tous les couples de couleurs possibles et la troisième traduit la distance spatiale entre les couleurs concernées.

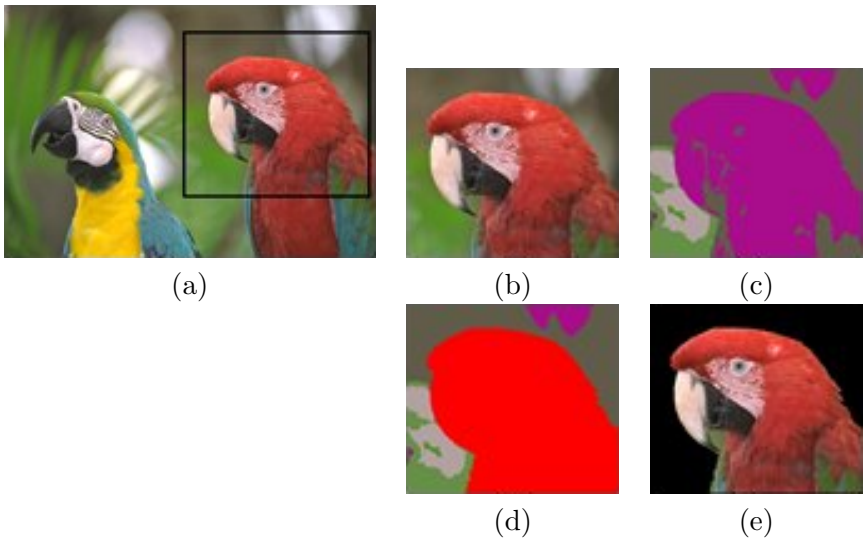
Toutefois, ces approches restent assez limitées à des requêtes bas-niveaux, et ne permet pas de manipuler des informations relatives aux objets présents dans l'image. En outre, elles sont assez sensibles aux conditions d'éclairage ou aux ombres portées sur les objets.

1.3.2 Partition par segmentation forte et caractéristiques d'objets

1.3.2.1 Notion de segmentation forte

Ce type de partition permet de découper l'image en régions qui correspondent exactement aux objets du monde réel. Il faut noter que la segmentation forte est un problème extrêmement complexe à résoudre. Lorsque les images à traiter sont particulièrement simples (par exemple un objet unique sur un fond uniforme), il est possible d'obtenir une segmentation forte. Mais dès que les images présentent de forte

variabilité (d'éclairage, mais aussi dans le type et le nombre d'objets représentés, etc...), la segmentation forte est pratiquement impossible à obtenir sans intervention de l'humain.



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

FIG. 1.14 – Exemple de sélection semi-automatique de zones d'intérêt du système IMALBUM.

Ainsi, le système IMALBUM (IDRISSI ET AL., 2004) demande tout d'abord à l'utilisateur de sélectionner dans l'image une zone d'intérêt (figure 1.14(a)) sur laquelle est lancée une segmentation couleur automatique (c). L'utilisateur regroupe alors à sa guise les régions issues de cette segmentation (d) afin de former un objet d'intérêt, qui sera indexé (e).

Un traitement similaire est proposé dans le système QBIC (FLICKNER ET AL., 1995).

1.3.2.2 Notion de forme

Si cette intervention de l'humain dans le processus de segmentation constitue une limitation évidente, ce type de partition permet néanmoins d'extraire des zones concernées des caractéristiques particulièrement robustes et discriminantes. Nous les appellerons caractéristiques d'objets car elles pré-supposent une segmentation forte et manipulent donc des informations liées aux objets réels. En pratique, ces caractéristiques permettent de décrire la forme des objets considérés.

Différentes méthodes pour caractériser la forme

Remarquons d'emblée que cette notion de forme ne fait pas l'objet d'une définition claire et consensuelle. Ainsi, de nombreuses approches de ce type existent, que nous ne détaillerons pas. Pour une présentation détaillée, nous renvoyons à l'article de Zhang et Lu (2004). D'une manière générale, on peut classer les descriptions de type forme en deux familles :

- celles qui décrivent les objets selon leur distributions spatiales de pixels sont qualifiées de caractéristiques basées régions.
- celles qui décrivent les objets selon leur contour externe sont qualifiées d'approches contours.

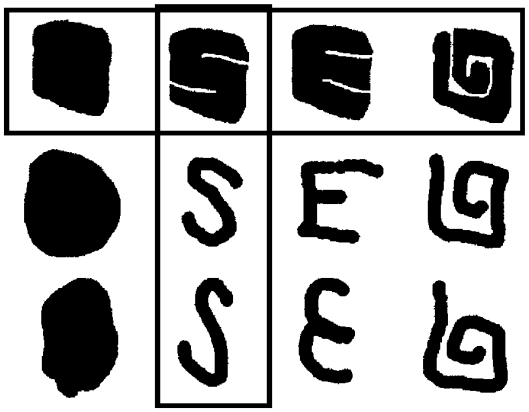


Fig. 1.15 – Exemple de similarité de formes d'après la région ou le contour. Reproduit de Zhang et Lu (2004).

La figure 1.15 illustre cette distinction. Considérons par exemple les objets de la première ligne. Puique leur distribution spatiale de pixels est semblable, ils seront jugés similaires par un descripteur de forme basé région. Toutefois, leur contours sont extrêmement variés. Au contraire, les objets de chaque colonne sont similaires du point de vue de leur contour, mais pas de leur distribution spatiale de pixels. Ainsi, si l'objet situé à l'intersection de la première ligne et de la deuxième colonne est posé en requête, un descripteur de forme basé région retournera uniquement les objets de la première ligne, mais pas ceux de la seconde colonne.

Il y a consensus (Zhang et Lu, 2004), en particulier au sein du comité MPEG-7 (Zaharia et Prêteux, 2004) sur le fait que les deux types de descripteurs de forme doivent être utilisés conjointement afin de fournir une réponse adéquate quant à la similarité de deux formes.

Pour chacune de ces approches (région ou contour), on peut ensuite distinguer deux sous-familles :

- celles qui décrivent les objets comme un tout (globales)

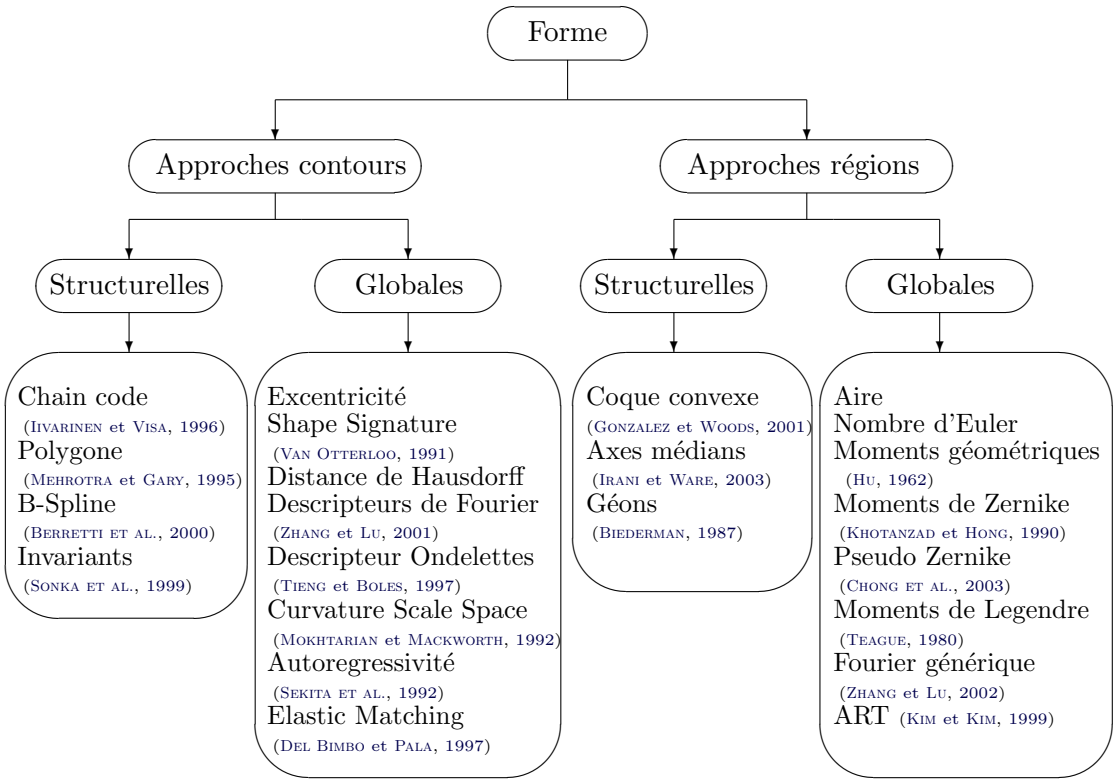


Fig. 1.16 – Classification des techniques de descriptions de formes 2D.

– celles qui décrivent les objets en les considérant comme un arrangement de sous-parties (structurelles)

La figure 1.16 récapitule ces différentes distinctions et cite, en outre, les principales méthodes et références associées.

Quelques propriétés des caractéristiques de formes

Devant la grande diversité des méthodes existantes pour caractériser la forme, il est utile de conserver à l'esprit un certain nombre de propriétés, que chaque méthode devrait respecter. En effet, dans une perspective d'indexation, lorsque nous modélisons la forme d'un objet par un descripteur, il est souhaitable que ce descripteur soit invariant à un certain nombre de transformations. Citons ici :

– translation
– changement d'échelle
– rotation

La figure 1.17 présente des exemples de ces transformations.

Ajoutons aussi une autre propriété, très importante lors du traitement d'images naturelles : la résistance de la description au bruit.

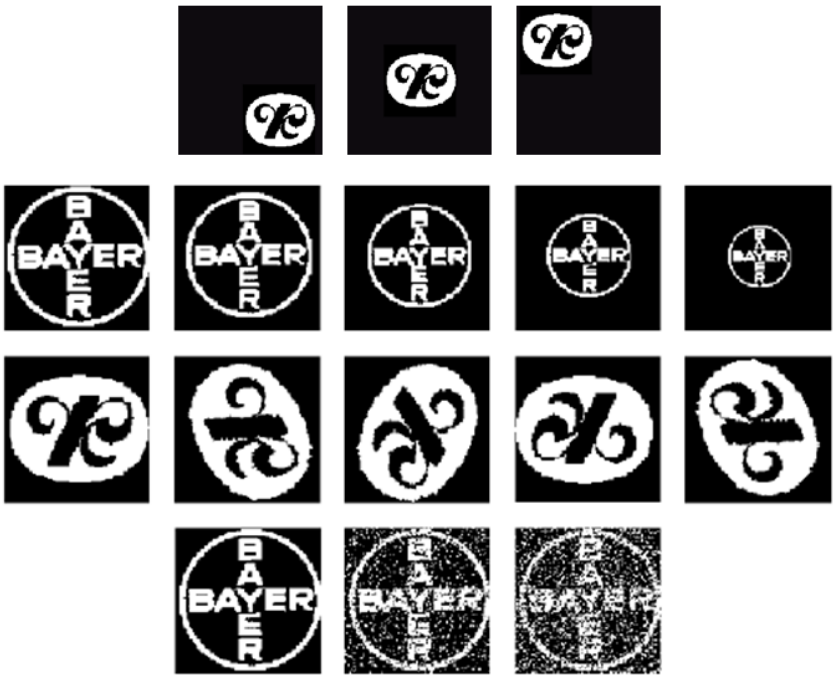


FIG. 1.17 – Exemples de transformations (en ligne) : translation, changement d'échelle, rotation et bruit (poivre et sel).

Nous allons maintenant présenter rapidement deux méthodes de description de forme, l'une orientée région et l'autre contour. Nous choisissons de mettre en exergue ces deux méthodes tout d'abord parce qu'elles ont été choisies par le comité MPEG-7 (ZAHARIA et PRÊTEUX, 2004) comme étant particulièrement robustes et discriminantes. En outre, comme nous utilisons ces deux méthodes plus loin dans nos travaux, nous jugeons pertinent de les introduire maintenant.

1.3.2.3 Description de forme basée région : méthode ART

ART (*Angular Radial Transform*) est un descripteur de forme robuste au changement d'échelle, à la translation et à la rotation. Il consiste à projeter l'objet à étudier sur une série de fonctions de base (KIM et KIM, 1999). Formellement, ART se définit comme une transformation 2D complexe, sur un disque unité, en coordonnées polaires. Les coefficients ART $F_{m,n}$ d'ordre m and n d'un objet f sont obtenus par la formule suivante :

$$F_{m,n} = \int_0^{2\pi}\int_0^1 \frac{1}{2\pi} V_{m,n}(\rho,\theta) f(\rho,\theta)\, \rho \,\mathrm{d}\rho \,\mathrm{d}\theta \tag{1.16}$$

$V_{m,n}$ représentent les fonctions de base ART, qui sont séparables dans les directions angulaires et radiales :

$$V_{m,n}(\rho,\theta) = A_m(\theta) R_n(\rho) \tag{1.17}$$

avec

$$A_m(\theta) = \tfrac{1}{2\pi} \mathrm{e}^{jm\theta} \quad R_n(\rho) = \begin{cases} 1 & n = 0 \\ 2\cos(\pi n \rho) & n \neq 0 \end{cases} \tag{1.18}$$

La fonction de base d'ART V_{mn} est une fonction complexe. La figure 1.18 représente la partie réelle de chaque fonction de base. Les parties imaginaires sont identiques aux parties réelles, avec une différence de phase.

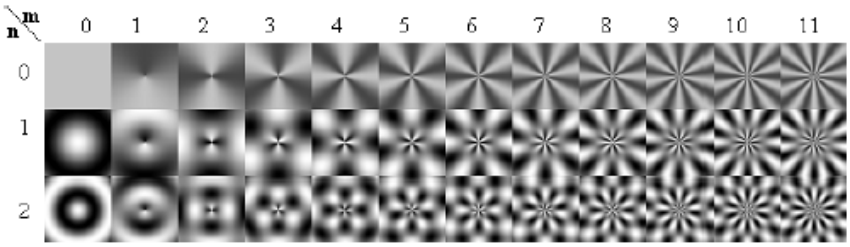


FIG. 1.18 – Parties réelles des fonctions de base ART V_{mn}.

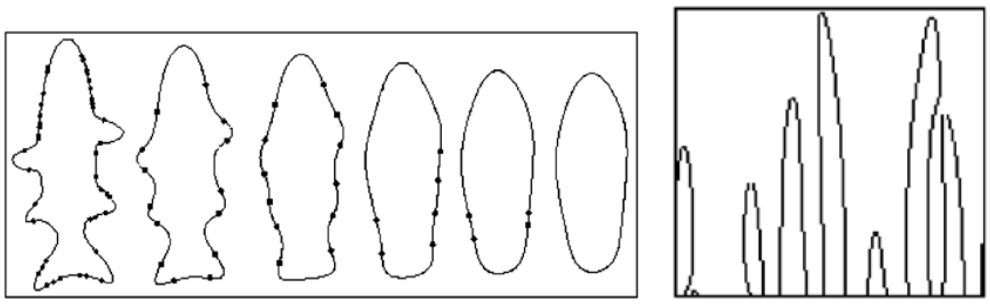
La transformation ART peut être considérée comme la projection d'un objet sur chaque fonction de base. Ainsi, pour un objet donné, le coefficient F_{nm} est d'autant plus grand que l'objet est semblable à la fonction de base V_{nm}.

Pour chaque objet, on aura n * m coefficients ART. La question se pose de savoir combien sont nécessaires à la description d'un objet. Le comité MPEG-7 (JEANNIN, 2001) propose de fixer n à 3 et m à 12.

1.3.2.4 Description de forme basée contour : méthode CSS

CSS (*Curvature Scale Space*), introduit par MOKHTARIAN et MACKWORTH (1992) est un descripteur de forme robuste au changement d'échelle, à la translation et à la rotation. Il caractérise le contour des objets. Plus précisément, la

représentation CSS d'un contour fermé se construit en repérant les positions des points d'inflexion du contour, alors que ce dernier subi une série de filtrages gaussiens passe-bas. Au fur et à mesure des itérations de filtrage, le contour devient de plus en plus lisse et les inflexions non significatives sont éliminées (figure 1.19(a)). On considère que les inflexions qui subsistent à la fin des filtrages sont des caractéristiques saillantes de l'objet étudié.



(a) Filtrages gaussiens successifs (b) Représentation CSS

FIG. 1.19 – Filtrages gaussiens successifs et représentation CSS associée d'un objet (adapté de MOKHTARIAN et MACKWORTH (1992)).

En pratique, les points d'inflexion d'un contour sont extraits en cherchant les *zero-crossings* de la courbure K locale. Cette dernière est calculée avec la formule suivante :

$$K(j) = \frac{x'(j)y''(j) - y'(j)x''(j)}{(x'^2 + y'^2)^{3/2}} \tag{1.19}$$

En repérant à chaque itération de filtrage, la position des points d'inflexion par leur abscisse curviligne, on obtient une représentation CSS du contour (figure 1.19(b)). Cette dernière consiste en une série de pics, qui représentent les différentes inflexions du contour. Plus un pic est haut, plus l'inflexion correspondante a résisté aux filtrages. On considère alors qu'elle caractérise bien l'objet.

1.3.3 Partition par segmentation faible et caractéristiques saillantes

Devant l'impossibilité d'obtenir une segmentation forte, c'est-à-dire de pouvoir disposer de primitives (régions, points) correspondant exactement aux objets du monde 3D, les approches d'indexation basées contenu se contentent généralement d'une segmentation dite faible. Typiquement, les pixels de l'image sont regroupés en régions homogènes en terme de couleur ou de texture. Eventuellement, la phase suivante consiste à sélectionner parmi ces régions celles qui sont les plus caractéristiques par rapport au reste de l'image. Un cas limite consiste à n'extraire que des pixels caractéristiques.

Les descripteurs extraits à partir de ces primitives régions ou points, sont qualifiés de caractéristiques saillantes parce qu'elles ont été jugées significatives par rapport au reste de l'image.

1.3.3.1 Approches empiriques

Par exemple, Carson et al. (2002) modélisent une image par un mélange de gaussiennes et utilisent le principe EM (*Expectation-Maximization*, Dempster et al. (1977)) pour estimer les paramètres de ce modèle. Cela conduit à une description de l'image à base de régions adjacentes (ou *blobs*), qui présentent des caractéristiques différentes en terme de couleur ou de texture.
La figure 1.20 présente un exemple de cette description. Notons que la sélection des régions saillantes se fait a posteriori, en ne conservant que les plus grandes (les pixels gris de la figure 1.20(b) n'ont ainsi pas été conservés). On remarque en outre que la segmentation ne parvient pas à regrouper les différentes parties du poisson du fait de la trop grande dissimilarité des caractéristiques couleur et texture correspondantes.

(a) Image originale (b) Image segmentée

Fig. 1.20 – Exemple de description de type *Blobworld*.

Sur le même principe, Wang et al. (2001) décrivent chaque pixel de l'image dans un espace à six dimensions (trois de couleurs et trois de texture) qui est ensuite partitionné en classes avec un algorithme de nuées dynamiques non supervisé.

Notons qu'il est possible, dans une certaine mesure, de corriger une partie des erreurs dues à la segmentation en utilisant une méthode de comparaison adaptée. Ainsi, Wang et al. (2001) propose la mesure IRM (*Integrated Region Measure*) qui compare une région de la requête avec plusieurs régions de chaque image testée. Toutefois, ce mécanisme ne permet pas de rectifier d'importantes erreurs de segmentation.

Une alternative est de ne faire aucune segmentation, et de ne traiter que des points d'intérêts. Ainsi, Tuytelaars et Gool (1999) proposent d'indexer une image par une liste d'invariants locaux, extraits en un certain nombre de points d'intérêt. Ces derniers sont extraits par un détecteur de coins proposé par Harris et Stephens (1988). Ce dernier consiste à utiliser la fonction d'auto-corrélation afin

de déterminer les sites de l'image où le signal change dans deux directions à la fois. En pratique, ils calculent une matrice A qui fait intervenir les dérivées d'ordre 1 (L_x et L_y) et 2 (L_x^2 et L_y^2).

$$A = G(\sigma) * \begin{pmatrix} L_x^2 & L_x L_y \\ L_x L_y & L_y^2 \end{pmatrix} \tag{1.20}$$

Les vecteurs propres de A associés à des valeurs propres non négligeables fournissent les points d'intérêt. La figure 1.21 présente deux exemples d'extraction de points d'intérêt par cette méthode.

(a) (b)

FIG. 1.21 – Extraction de points d'intérêt de type coins par la méthode de HARRIS et STEPHENS (1988).

De la même manière, ROS ET AL. (2005) proposent d'indexer une image par une liste de masques 3x3, centrés en des points caractéristiques. Ces derniers correspondent cette fois à des zones de fort contraste, et sont extraits par un dispositif approprié.

1.3.3.2 Approches psycho-visuelles

La sélection de zones saillantes dans une image demeure un problème de recherche ouvert, qui s'inscrit dans un contexte pluridisciplinaire, touchant à la psychologie, les neurosciences ou encore à la biologie. Ainsi, un certain nombre d'approches, dites psycho-visuelles, ont tenté de modéliser le fonctionnement du système visuel humain, afin de pouvoir extraire les zones saillantes de l'image.

KOCH et ULLMAN (1985) ont défini le premier modèle biologiquement plausible de l'attention visuelle. Ce dernier stipule que le système visuel humain extrait un certain nombre de primitives de l'image (intensité, contraste, orientations) et crée pour chacune d'elle un carte de caractérisation qui rend compte de la topologie de la source. Toutes ces cartes sont ensuites combinées pour créer une carte unique de

saillance, qui présente, pour chaque pixel, son "pouvoir attracteur" relativement à l'image.

Itti et al. (1998) ont ensuite proposé, dans cette continuité, le modèle de l'attention visuelle qui fait aujourd'hui référence. Les primitives bas-niveaux utilisées concernent l'intensité lumineuse, la couleur, ainsi que les orientations de filtres de Gabor.

Nous ne détaillerons pas ici ces modèles, car nous y reviendrons largement dessus dans la partie 2.

1.3.4 Caractéristiques de relations spatiales

Les méthodes que nous avons vues jusqu'à maintenant permettent d'extraire de l'image plusieurs primitives (points, régions) et de stocker des descriptions liées intrinsèquement à celles-ci comme la couleur, la forme, etc. Mais il est également possible de considérer l'organisation spatiale des différentes primitives comme une description en tant que telle. Il est évident qu'une telle description constitue un niveau intermédiaire entre les données bas-niveaux brutes et l'interprétation des images, et que ce niveau peut se révéler très expressif. Ainsi, lorsqu'un utilisateur cherche une image représentant un objet complexe (par exemple : une voiture), la reconnaissance de chacune de ses parties (roues, carrosserie...) est une étape indispensable. Toutefois, vérifier en sus que chacune de ces parties respecte certaines contraintes spatiales (les roues en-dessous de la carrosserie, etc...) est particulièrement important.

Partant de cette constatation, Smith et Chang (1999) stockent pour chaque région extraite sa position, sa taille et des caractéristiques bas-niveau. La correspondance entre images est ensuite basée sur les chaînes 2D introduites par Chang et Hsu (1992). D'une manière générale, les chaînes 2D permettent de modéliser des arrangements spatiaux de formes simples, qui ne se recouvrent pas. En outre, la distance est peu robuste aux changements locaux.

Dans des domaines plus spécifiques, comme les images médicales à rayons X, Petrakis et Faloutsos (1997) modélisent des régions extraites par des noeuds dans un graphe. Les arêtes de ce derniers encodent alors les relations spatiales entre les régions. Toujours dans un domaine particulier, Forsyth et Fleck (1999) parviennent à reconnaître des personnes nues dans des images, en assemblant des régions de couleurs chairs selon des configurations pré-établies, correspondant à différentes postures humaines (figure 1.22).

Toutefois, ces exemples intègrent un certain nombre de procédures ad-hoc dans leur fonctionnement, ce qui rend difficile leur généralisation à des domaines plus larges ou moins contraints.

En outre, le facteur limitant de ces approches vient encore une fois de la segmentation, puisqu'il faut être certain que l'on extrait des structures significatives, et

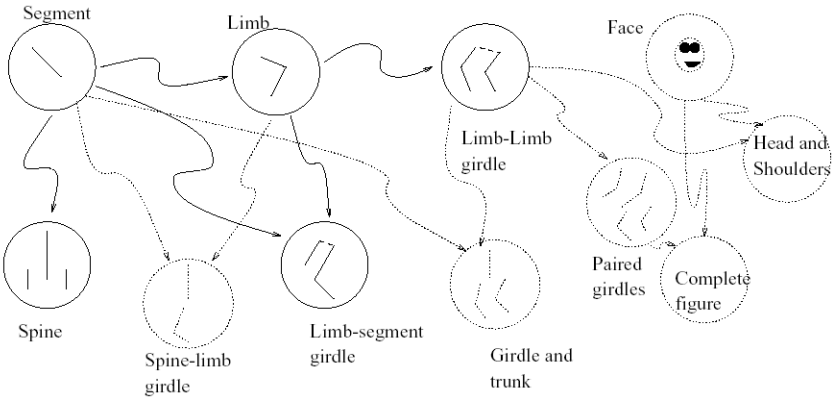


FIG. 1.22 – Processus d'extraction de régions pour la reconnaissance de corps (FORSYTH et FLECK, 1999).

non pas des arrangements accidentels de régions, sans rapport aucun avec les objets 3D du monde réel.

1.3.5 Discussion sur l'extraction de caractéristiques

Il parait évident que la partition la plus efficace pour une indexation est la segmentation version forte, puisqu'elle permet de manipuler des zones de l'image qui correspondent aux objets du monde 3D. En outre, on dispose à ce niveau de description de tout un arsenal de descripteurs de type forme, particulièrement discriminant.

Néanmoins, cette segmentation est impossible à réaliser dans le cas général. Un domaine de recherche particulièrement critique concerne donc la version faible de la segmentation. Plus celle-ci est grossière (comme dans l'exemple de Blobworld), et moins on pourra utiliser des descripteurs de type objet comme la forme, car ces derniers manipulent alors des régions sans réelle cohérence sémantique (artefact intégré dans la région, ou fusion de deux objets en un seul par exemple).

Si de nombreuses méthodes de segmentation ont vu le jour, il ne faut pas oublier qu'elles restent extrêmement limitées lorsqu'elles s'appuient sur des descriptions de type couleur ou texture. En effet, ces dernières sont trop sensibles aux conditions de prise de vues et à d'autres artefacts. En outre, de telles méthodes restent limitées pour extraire des objets du monde 3D, souvent composés de différentes parties, avec pour chacune des descriptions bas-niveaux très différentes.

Enfin, une voie très explorée à l'heure actuelle consiste... à ne faire aucune segmentation. Les descripteurs globaux permettent ainsi d'effectuer un certain nombre de traitements, même si leur utilisation est particulièrement limitée. Par exemple, VAILAYA ET AL. (1998) proposent un système de classification d'images (intérieur, extérieur...) basé sur des histogrammes couleur. Plus prometteur, SCHNEIDERMAN

et KANADE (2000) utilisent des histogrammes de descripteurs ondelettes, pour rechercher des visages ou des voitures à partir d'un apprentissage.

1.4 Requête

Une fois que des caractéristiques ont été extraites de l'image, deux cas sont maintenant envisageables concernant la requête de l'utilisateur au système :

- les caractéristiques peuvent être interprétées directement par le système, qui est alors en mesure de dériver une description symbolique de chacune des images de la base. L'interrogation à la base se fait alors via des mots, ou *labels sémantiques*.
- Les caractéristiques extraites constituent en elles-mêmes l'index de chacune des images. Dans ce cas, l'utilisateur soumet une requête au système en fournissant lui-même une liste de caractéristiques à rechercher. Ceci peut en particulier se faire en soumettent une image d'exemple.

Nous détaillons maintenant ces deux types d'utilisation.

1.4.1 Interprétation

Il s'agit ici de dériver une interprétation sémantique de l'image, à partir des caractéristiques extraites. Une approche très répandue consiste à effectuer un apprentissage bayesien.

1.4.1.1 Principe de l'apprentissage bayesien

Notons $\mathbf{C}$ un vecteur de caractéristiques extraites d'une image, et s une interprétation sémantique. Par exemple, s peut désigner l'interprétation *l'image contient une voiture*. Il s'agit alors de trouver l'image qui rend maximale la probabilité conditionnelle $P(s/\mathbf{C})$. L'application de la formule de Bayes donne :

$$P(s/\mathbf{C}) = \frac{P(\mathbf{C}/s) \cdot P(s)}{P(\mathbf{C})} \tag{1.21}$$

La quantité $P(\mathbf{C}/s)$ est appelée probabilité a posteriori du descripteur $\mathbf{C}$. Elle peut être évaluée lors d'une phase d'apprentissage, en soumettant au système un ensemble d'images dont l'interprétation a été effectuée manuellement. La quantité $P(\mathbf{C})$ est appelée probabilité a priori du descripteur $\mathbf{C}$. Elle peut également être évaluée lors de l'apprentissage. La quantité $P(s)$ est plus complexe à évaluer, mais son importance n'est pas primordiale.

Ainsi, ayant extrait les caractéristiques $\mathbf{C}$ d'une image, et étant données $P(\mathbf{C}/s)$ et $P(\mathbf{C})$ à l'issue d'un apprentissage, il est possible d'en déduire $P(s/\mathbf{C})$ via l'équation 1.21. La figure 1.23 illustre le mécanisme.

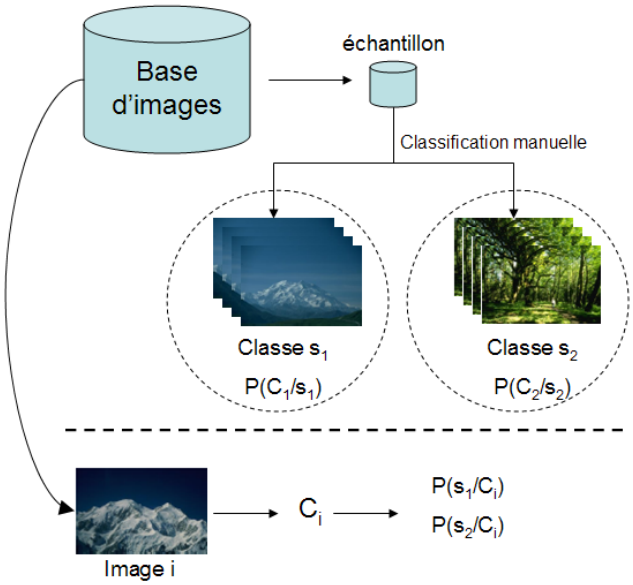


FIG. 1.23 – Principe de l'apprentissage bayesien.

Si l'on dispose de plusieurs interprétations potentielles, on retient généralement l'interprétation s_i qui maximise la probabilité a posteriori $P(s_i/\mathbf{C})$. On parle du critère de Bayes, ou encore de choix MAP (Maximim A Posteriori).

1.4.1.2 Exemple de résultats

L'apprentissage bayesien a beaucoup été utilisé ces dernières années et a conduit à de nombreux résultats de qualité. Leur efficacité dépend évidemment du ou des descripteur(s) utilisés pour l'apprentissage. Ainsi, SCHNEIDERMAN et KANADE (2000) proposent un sytème capable de reconnaître dans une image des visages ou des voitures, bien que ces classes recouvrent une large gamme de cas. Pour ceci, les auteurs utilisent comme descripteurs des histogrammes d'ondelettes. Notons toutefois que le système nécessite un apprentissage par point de vue (face, profil et trois-quart). La figure 1.24 présente des exemples de résultats pour les visages.

Dans une approche similaire, FERGUS ET AL. (2003) utilisent des masques invariants à l'échelle pour caractériser des voisinages de points d'intérêt. VAILAYA ET AL. (1998) proposent une série de classificateurs binaires, à utiliser séquentiellement. Chacun des classificateurs est basé sur des caractéristiques d'histogrammes de couleur. Ainsi, un tel système permet par exemple de catégoriser dans un premier temps des images selon deux classes : extérieur et intérieur. Puis, dans un deuxième temps,

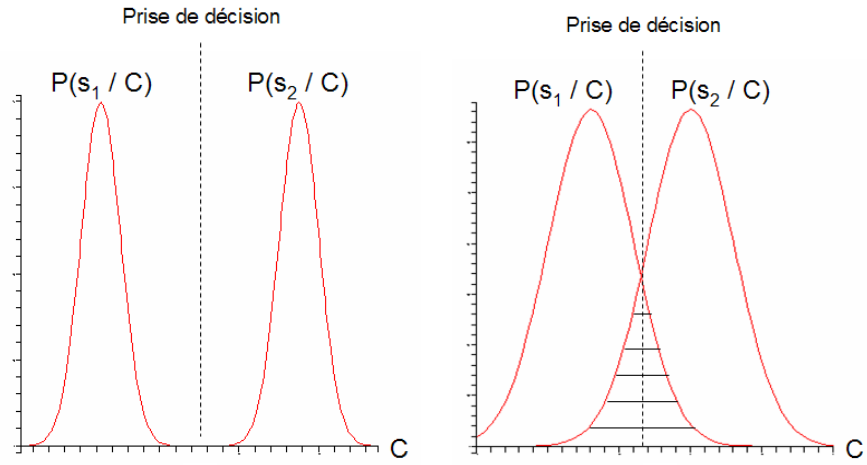
FIG. 1.24 – Exemple de reconnaissance de visages d'après SCHNEIDERMAN et KANADE (2000).

les images d'extérieur peuvent être passées à un deuxième classificateur qui les séparera en deux sous-classes : paysages naturels d'une part et "artificiels" (villes) d'autre part.

1.4.1.3 Limites des approches par apprentissage

La limitation fondamentale de ce type d'approches provient du pouvoir discriminant des caractéristiques **C** employées. En effet, si elles ne sont pas spécifiques à l'interprétation s recherchée, une autre interprétation s_2 pourra être associée à **C**. La figure 1.25 illustre le problème. En (a), les distributions de probabilité $P(s_1/C)$ et $P(s_2/C)$ sont bien distinctes. Le critère de choix à formuler peut donc être efficace et séparer sans ambiguité les deux distributions. En (b) en revanche, du fait du recouvrement partiel des distributions, le critère de choix aboutira nécessairement à une mauvaise classification de certaines occurences (rayées).

Cette limitation est une conséquence directe du fossé sémantique : une série de descripteurs ne suffit pas à elle seule à définir une interprétation fiable. D'autres informations doivent être intégrées afin de lever les ambiguités éventuelles (prise en



(a) (b)

FIG. 1.25 – Limitation du principe d'apprentissage.

compte du contexte, etc...).

En outre, ces approches, basées directement pixels, sont extrêmement côuteuses en temps de calcul. La phase d'apprentissage est un processus statistique et nécessite donc une collection importante de données, même si de nombreux travaux ont permis d'accélérer les procédures (LOWE, 2004; TORRALBA ET AL., 2004). En outre, il faut bien garder à l'esprit que ce type de système aboutit à une reconnaissance statique d'objets : à l'issue d'un processus d'apprentissage, l'utilisateur dispose d'une boîte noire qui, étant donné une image, concluera à la présence ou à l'absence d'un objet donné dans cette image. Aucun autre type de requête ne peut être formulé et aucune autre interaction avec l'utilisateur ne peut avoir lieu, par exemple, pour affiner le résultat.

1.4.2 Mesure de similarité

Il s'agit cette fois de comparer une signature fournie par l'utilisateur avec celles des images d'une base. Cela permet, par exemple, de retourner à l'utilisateur les images d'une base qui ont été jugées similaires à une image requête. On peut également utiliser le mécanisme de similarité pour organiser la base en groupes d'images.

1.4.2.1 Similarités attentive et pré-attentive

Il convient de distinguer à ce stade deux types de similarité d'images : d'une part la similarité pré-attentive, qui juge l'image dans sa globalité et d'autre part la similarité attentive, qui intervient une fois que l'attention de l'utilisateur a été focalisée. En effet, il est prouvé que la perception visuelle humaine opère selon ces

deux phases (Levy-Schoen, 1976). La première (très rapide, inférieure à 100 ms) fait intervenir des mécanismes globaux sur l'image, et correspond à des traitements qui ont lieu avant le cortex dans le cerveau humain. Puis, dans une deuxième phase, le regard opère un certain nombre de fixations sur l'image, et étudie, pour chacune d'elles, un proche voisinage.

Par exemple, la figure 1.26 montre deux images qui sont jugées très similaires, voire identiques, mais en retournant celles-ci, on constate qu'elle ne le sont plus du tout ! En effet, lorsque l'on regarde ces images à l'endroit, notre cerveau sait qu'il a affaire à des visages, et opèrent donc un certain nombre de traitements spécifiques pour comparer les visages en des points caractéristiques (les yeux, le nez par exemple). A cet instant, toute légère différence en ces points caractéristiques entraine un forte dissimilarité entre les images. Au contraire, lorsqu'elles sont perçues à l'envers, c'est-à-dire dans une configuration inhabituelle, le cerveau ne parvient plus à lancer ces divers traitements, et se contente d'une similarité globale, en ne comparant plus des *visages* mais des *images* à travers certaines caractéristiques.

Fig. 1.26 – Exemple de similarités pré-attentive et attentive.

Il est toutefois important de noter que cette distinction entre vision attentive et pré-attentive est plus subtile qu'il n'y parait. En effet, des travaux ont montré que même lors d'une phase de perception pré-attentive, une part d'interprétation peut déjà être à l'oeuvre. Par exemple, des catégorisations d'images (de type paysage de villes / naturels) peuvent se faire en moins de 100 ms par l'être humain (Oliva, 2005).

Par la suite, nous ne ferons plus cette distinction pré-attentive / attentive dans la similarité. Au contraire, nous la considèrerons sous un aspect plus fonctionnel.

1.4.2.2 La similarité par une fonction distance : axiomatique

Il est généralement admis que la similarité $S(R, I)$ entre une image requête R et une image tierce I est le résultat de la composition d'une fonction g strictement décroissante et d'une distance d entre les caractéristiques $\mathbf{C}_R$ et $\mathbf{C}_I$ de R et I respectivement, soit :

$$S(R, I) = g \circ d(\mathbf{C}_R, \mathbf{C}_I) \tag{1.22}$$

Ainsi, deux images seront jugées d'autant plus similaires que la distance entre leurs caractéristiques aura été faible.

Mathématiquement, une distance d se définit comme une fonction à valeur dans $[0\,;1]$ et qui vérifie les trois propriétés suivantes :

$$d(\mathbf{C}_1, \mathbf{C}_2) + d(\mathbf{C}_2, \mathbf{C}_3) \geq d(\mathbf{C}_1, \mathbf{C}_3) \tag{1.23}$$
$$d(\mathbf{C}_1, \mathbf{C}_2) = d(\mathbf{C}_2, \mathbf{C}_1) \tag{1.24}$$
$$d(\mathbf{C}_1, \mathbf{C}_2) = 0 \Leftrightarrow \mathbf{C}_1 = \mathbf{C}_2 \tag{1.25}$$

La condition 1.23 est connue sous le nom d'inégalités triangulaire, la 1.24 est appelée symétrie. Nous reviendrons sur la pertinence de ces axiomes lorsque nous évoquerons la similarité de Tversky.

La plupart des systèmes existant (pour ne pas dire l'immense majorité), utilisent une distance pour mesurer une similarité. Notons qu'une fonction distance se définit différemment selon qu'elle opère sur des vecteurs de caractéristiques ou sur des descripteurs de type accumulatif, tels que les histogrammes. Nous allons maintenant présenter les distances les plus couramment utilisées.

1.4.2.3 Exemples de fonctions distance entre vecteurs

Concernant les vecteurs de caractéristiques, on définit la famille L_p de distances de Minkowski, entre deux vecteurs $\mathbf{C}_1$ et $\mathbf{C}_2$ par :

$$L_p(\mathbf{C}_1, \mathbf{C}_2) = \left(\sum_{i=0}^{n} |\mathrm{C}_{1,i} - \mathrm{C}_{2,i}|^p \right)^{\frac{1}{p}} \tag{1.26}$$

Où $\mathrm{C}_{k,i}$ désigne la i-ème composante du vecteur $\mathbf{C}_k$. En faisant varier la valeur de p,on obtient ainsi différentes fonctions de distance.

Distance de Minkowski

Avec $p = 1$, on obtient $L_1(\mathbf{C}_1, \mathbf{C}_2) = \sum_{i=0}^{n} |\mathrm{C}_{1,i} - \mathrm{C}_{2,i}|$. Cette distance est aussi connue sous le nom de *city-block* ou encore Manhattan.

C'est cette distance qui est recommandée par le comité MPEG-7 (Jeannin, 2001) pour comparer deux formes décrite avec la méthode ART de Kim et Kim (1999) (voir section 1.3.2.3). Ainsi, pour deux objets O_1 et O_2, décrits respectivement par une série de m coefficients $F_{1,i}$ et $F_{2,i}$, on peut évaluer une mesure de dissemblance avec la formule :

$$L_1(\mathbf{O}_1, \mathbf{O}_2) = \sum_{i=0}^{m} |\mathrm{F}_{1,i} - \mathrm{F}_{2,i}| \tag{1.27}$$

Distance euclidienne

Avec $p = 2$, on obtient $L_2(\mathbf{C}_1, \mathbf{C}_2) = \sqrt{\sum_{i=0}^{n} |\mathrm{C}_{1,i} - \mathrm{C}_{2,i}|^2}$. C'est le cas le plus fréquemment rencontré.

Distance uniforme

En faisant tendre p vers l'infini, on peut montrer que l'on obtient une nouvelle mesure limite (au sens d'une fonction) qui est elle-même une distance. On la note L_∞ telle que : $L_\infty = \max_{i=0}^{n} |\mathrm{C}_{1,i} - \mathrm{C}_{2,i}|$.

1.4.2.4 Exemples de fonctions distance entre fonctions accumulatives

Distance classique entre histogrammes

Swain et Ballard (1991) ont défini une distance $d(\mathbf{H}_1, \mathbf{H}_2)$ entre deux histogrammes $\mathbf{H}_1$ et $\mathbf{H}_2$ par la notion d'intersection, définie comme suit :

$$d(\mathbf{H}_1, \mathbf{H}_2) = \sum_{j=1}^{n} \min(H_{1,j}, H_{2,j}) \tag{1.28}$$

où $\mathrm{H}_{k,j}$ désigne la j-ème composante de l'histogramme $\mathbf{H}_k$ et n le nombre de discrétisations de chaque histogramme. Les auteurs ont en outre montré que si toutes les images concernées possédaient le même nombre de pixels (c'est-à-dire $\sum_{j=1}^{n} H_{1,j}$ constant pour toutes les images), alors cette mesure possède les mêmes propriétés ordinales qu'une mesure L_1. Dans ce cas, on peut normaliser la mesure de façon à obtenir :

$$d(\mathbf{H}_1, \mathbf{H}_2) = \frac{\sum_{j=1}^{n} \min(H_{1,j}, H_{2,j})}{\sum_{j=1}^{n} H_{2,j}} \tag{1.29}$$

Distance entre histogrammes par forme quadratique

Hafner et al. (1995) proposent quant à eux une mesure pondérée qui permet d'intégrer la notion de similarité entre deux couleurs.

$$d(\mathbf{H}_1, \mathbf{H}_2) = \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} \left(H_{1,i} - H_{2,j}\right)^2} \tag{1.30}$$

où $\mathrm{H}_{k,j}$ désigne la j-ème composante de l'histogramme $\mathbf{H}_k$ et a_{ij} la similarité entre les deux couleurs i et j. Cette dernière est calculée d'après la distance euclidienne d_{ij} entre les deux couleurs, via la formule :

$$a_{ij} = 1 - \frac{d_{ij}}{\max_{i,j} d_{ij}} \tag{1.31}$$

Distance statistique entre histogrammes

Stricker et Dimai (1996) considèrent un histogramme comme une densité de probabilité et proposent alors de calculer pour chaque canal (ie : chaque axe de l'espace de représentation) la moyenne, la variance et le moment statistique d'ordre 3, afin de caractériser la distribution statistique. La distance entre deux histogrammes $\mathbf{H}_1$ et $\mathbf{H}_2$ est alors obtenue par :

$$d(\mathbf{H}_1, \mathbf{H}_2) = \sum_i W_{i,1} \left|\mu_i^{\mathbf{H}_1} - \mu_i^{\mathbf{H}_2}\right| + W_{i,2} \left|\sigma_i^{\mathbf{H}_1} - \sigma_i^{\mathbf{H}_2}\right| + W_{i,3} \left|m_i^{\mathbf{H}_1} - m_i^{\mathbf{H}_2}\right| \tag{1.32}$$

où $\mu_i^{\mathbf{H}_1}$, $\sigma_i^{\mathbf{H}_1}$ et $m_i^{\mathbf{H}_1}$ représentent respectivement la moyenne, la variance et le moment statistique d'ordre 3 du canal i de l'histogramme $\mathbf{H}_1$. Les $W_{i,j}$ représentent les pondérations associées à chaque terme.

Autres fonctions accumulatives

Il existe évidemment d'autres fonctions accumulatives que les histogrammes. Par exemple, nous avons parlé à la section 1.3.2.4 de la représentation CSS d'un contour. La principale difficulté de ce type de représentation provient de la méthode de comparaison à fournir. Dans le cas de deux représentations CSS d'objet O_1 et O_2, le

comité MPEG (Jeannin, 2001) propose d'utiliser une mesure L_2 entre les paires de pics qui se correspondent. Une pénalité est ajoutée pour chaque pic qui n'est pas mis en correspondance. Ainsi :

$$D_{CSS}(S_1, S_2) = \sum_1 \left((x_{S_1,i} - x_{S_2,j})^2 + (y_{S_1,i} - y_{S_2,j})^2 \right) + \sum_2 y_i^2 \tag{1.33}$$

où $\sum_1$ représente la sommation sur tous les pics mis en correspondance et $\sum_2$ la sommation sur tous les pics non appariés.

1.4.2.5 Autre similarité : le modèle de Tversky

Toutes les mesures que nous venons de voir vérifient les axiomes d'une fonction distance. Ainsi, la quasi-totalité des systèmes d'indexation existants utilisent une distance pour établir leur mesure de similarité.

Néanmoins, un certain nombre de travaux montrent l'inadéquation des mesures de distances par rapport au jugement humain de similarité. Il est étonnant d'ailleurs de constater que, bien que ces résultats soient connus de la communauté indexation, très peu de travaux s'attèlent à une réflexion sur la notion même de similarité (Santini, 2001).

Nous allons maintenant présenter de façon succinte quelques résultats qui remettent en cause l'axiomatique d'une distance par rapport au jugement humain de similarité.

Propriété de symétrie

Ainsi, et de façon assez surprenante au premier abord, Tversky (1977) montre que notre jugement de similarité n'est pas symétrique. En particulier la notion de modèle, ou de prototype joue un rôle primordial en la matière, puisqu'un objet "réel" est toujours jugé plus similaire à son modèle que le modèle ne l'est vis-à-vis de l'objet. La figure 1.27 montre un exemple d'un tel cas : la similarité de la figure de droite à celle de gauche est jugée plus importante que celle de la figure de gauche à celle de droite. La figure est gauche est vue comme un modèle de celle de droite.

Propriété d'inégalité triangulaire

Dans la même optique, Tversky et Gati (1982) montrent que l'intégalité triangulaire n'est pas non plus respectée pour le jugement humain de similarité d'images. Par exemple, l'image A de la figure 1.28 est jugée similaire (faible distance) à l'image B, grâce à la forme et à la couleur commune des deux objets qui les composent. De

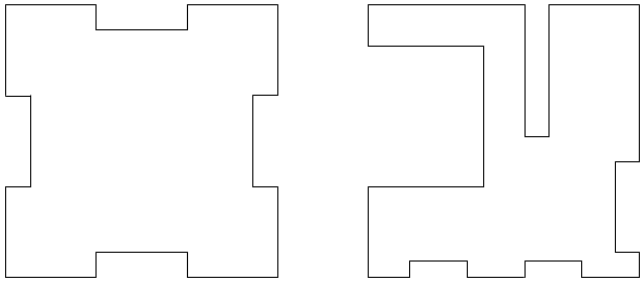


FIG. 1.27 – Exemple de cas où la règle de symmétrie n'est pas respectée (TVERSKY, 1977).

la même manière, l'image B est jugée similaire à l'image C puisqu'elles contiennent toutes deux des fruits. Cependant, l'image A est jugée très dissimilaire (forte distance) à l'image C, et on peut écrire, en notant $\delta(X, Y)$ la notion humaine de distance entre deux images X et Y :

$$\delta(A, C) > \delta(A, B) + \delta(B, C) \tag{1.34}$$

Dans ce cas, la notion humaine de distance δ ne respecte pas l'axiome de l'inégalité triangulaire.

A B C

FIG. 1.28 – Exemple de cas où l'inégalité triangulaire n'est pas respectée.

Le modèle de contraste de Tversky

Tous ces résultats ont ainsi conduit Tversky à proposer une autre mesure de similarité, plus en accord avec la perception humaine. Cette mesure est connue sous le nom des *modèles de contrastes* (TVERSKY, 1977). Dans le cas où l'on cherche à comparer une image I d'une base avec une image requête R, rappelons que l'on note $\mathbf{C}_R$ et $\mathbf{C}_I$ les caractéristiques extraites de R et I respectivement. Tversky propose que la similarité $S(R, I)$ entre R et I ait les propriétés suivantes :

– $S(R, I)$ est d'autant plus importante que R et I ont de caractéristiques communes

- $S(R, I)$ est d'autant moins importante que R possède de caractéristiques propres, que ne partage pas I
- $S(R, I)$ est d'autant moins importante que I possède des caractéristiques propres, que ne partage pas R

En notant alors $\mathbf{C}_R - \mathbf{C}_I$ les caractéristiques propres de R et $\mathbf{C}_I - \mathbf{C}_R$ celles de I, et $\mathbf{C}_R \cap \mathbf{C}_I$ celles communes à R et I, Tversky propose :

$$S(R, I) = f(\mathbf{C}_R \cap \mathbf{C}_I) - \alpha f(\mathbf{C}_R - \mathbf{C}_I) - \beta f(\mathbf{C}_I - \mathbf{C}_R) \tag{1.35}$$

On remarque par exemple tout de suite que dans le cas général ($\alpha \neq \beta$), cette mesure n'est pas symétrique.

Notons que cette mesure s'applique à des caractéristiques de prédicats logiques et que son application à des données numériques n'est pas sans poser quelques problèmes. Santini et Jain (1997) ont proposé une telle extension, grâce à l'utilisation de logiques floues.

1.5 Et après ? Evaluation des résultats

1.5.1 Evaluation globale

Il existe plusieurs moyens d'évaluer quantitativement les performances d'un système d'indexation. Les plus simples sont sans doute les deux suivants :

- le First Tier (FT), proposé par (Osada et al., 2001)
- le Second Tiers (ST), ou le *Score Bull-Eye* (SBE), retenu comme critère d'évaluation de tous les descripteurs de forme considérés dans MPEG-7.

Ces deux mesures sont définies par :

$$SBE = \frac{Ncor(2Q)}{Q} \qquad FT = \frac{Ncor(Q)}{Q} \tag{1.36}$$

où $Ncor(x)$ est le nombre de résultats corrects (i.e. image appartenant à la même catégorie que celle de la requête) parmi les x premiers résultats retrouvés et Q est le nombre d'éléments de la catégorie correspondant à la requête. Si le SBE est relativement permissif, admettant une marge d'erreur de 50 % pour retrouver l'ensemble des modèles recherchés, le FT est beaucoup plus drastique, un FT de 100% n'autorisant aucune erreur.

1.5.2 Evaluation fine

Pour une analyse plus fine des résultats, nous disposons des mesures de rappel (*recall*) et précision. Plus précisément :

- le rappel représente le nombre de réponses positives obtenues au rang k sur le nombre d'éléments de la classe de l'objet requête
- la précision représente le nombre de réponses positives au rang k sur le nombre total de réponses obtenues.

Ainsi :

$$Recall(k) = \frac{Ncor(k)}{Q} \qquad Precision(K) = \frac{Ncor(k)}{k} \tag{1.37}$$

Le Recall mesure la capacité du système à retrouver toutes les images pertinentes, alors que la précision mesure la capacité du système à retrouver seulement les images pertinentes. Ils sont donc évidemment complémentaires.

Concrètement, ils sont calculés pour tous les rangs k de la base et affichés sous forme de courbes : la courbe de rappel, la courbe de précision et la courbe de rappel/précision. Cette dernière est couramment utilisée car elle donne une représentation visuelle synthétiques des résultats.

1.5.3 Discussion

Notons que les FT et les SBE sont liées aux courbes de rappel. La valeur du FT correspond à la valeur de la courbe de rappel à l'abscise du cardinal de la classe et respectivement à deux fois le cardinal de la classe pour le SBE.

Il est important de constater la part de subjectivité qui demeure lors de l'utilisation de ces mesures. En effet, toutes reposent sur la notion de catégories ou classes. On considère que les images de la base peuvent être regroupées en familles, et que le système d'indexation doit pouvoir être capable de retrouver ces dernières. Toutefois, un tel regroupement n'est jamais absolu et dépend en grande partie de ce que souhaite manipuler l'utilisateur.

1.6 Conclusion et discussion sur notre contribution

Dans ce chapitre, nous avons présenté les différentes étapes à mettre en oeuvre dans un système d'indexation d'images. Loin de chercher à dresser une liste exhaustive des méthodes existantes, nous avions plutôt pour but de replacer chacune des étapes dans une perspective plus large : celle de l'indexation. Ainsi, nous n'avons pas listé toutes les méthodes de segmentations existantes, mais nous avons plutôt

présenté les résultats attendus et les problèmes rencontrés par les méthodes de segmentation en règle générale.

Cette vision nous a permis de tirer trois conclusions majeures.

L'extraction de caractéristiques est au coeur de l'indexation

Tout d'abord, une étape essentielle de l'indexation doit consister en une extraction de caractéristiques saillantes de l'image. En effet, les caractéristiques globales comme les histogrammes, bien que pertinentes, ne possèdent pas de pouvoir descriptif suffisamment fort pour pouvoir à elles seules, assurer l'indexation d'une image. Or, l'extraction de zones saillantes renvoie à des problématiques de vision par ordinateur extrêmement complexes et non résolues à l'heure actuelle. En outre, dans une perspective d'indexation, la notion de saillance est difficile à définir, puisqu'il s'agit de savoir si telle ou telle zone de l'image est saillante *a priori* c'est à dire avant toute requête, donc avant même de savoir ce que l'on cherche.
Quoi qu'il en soit, la question de l'extraction de zones saillantes reste une problématique centrale lors de l'indexation. Deux approches peuvent être envisagées, selon le type de primitives manipulées. La première à avoir été utilisée historiquement consiste à extraire des régions d'intérêt par une segmentation (couleur, texture ou extraction de contours). L'avantage de cette approche provient du fait qu'une fois que l'on dispose de zones d'intérêt, tout un arsenal de descripteurs de type forme, discriminants et expressifs, est disponible. La principale limitation, encore une fois, provient du fait qu'aucune méthode de segmentation n'est robuste, dans un univers non contraint.
C'est pourquoi une alternative consiste à ne pas faire de segmentation et à se contenter de primitives de type pixel. Bien qu'un certain nombre de résultats aient été obtenus avec cette méthode, la principale limitation a trait au faible pouvoir discriminant des descripteurs alors utilisés.
Nous avons donc proposé dans cette thèse une alternative au processus de segmentation (section 2). Sans chercher une méthode infaillible, nous proposons une segmentation multi-niveaux, qui permet d'indexer plusieurs zones d'intérêt dans chaque image. Nous verrons que cette méthode permet une indexation correcte, même si des erreurs de segmentations apparaissent (section 3.5.1).

La requête doit être pensée en terme d'usage

Une autre question centrale dans l'indexation concerne le processus de requête. Très peu de travaux proposent une réflexion sur les usages liés à un système d'indexation. Ainsi, la quasi-totalité des systèmes existants repose sur la désormais classique requête-par-l'exemple, même si sa pertinence en terme de souplesse d'utilisation n'est pas prouvée. Encore une fois, rien n'est prévu dans ce cas, pour que le système puisse déduire de l'exemple ce que cherche réellement l'utilisateur.

Pour la suite de nos travaux, nous nous sommes concentrés sur un cas d'utilisation : la recherche par l'utilisateur d'images représentant des *objets*. A ce titre, nous avons illustré notamment par les travaux de Tversky, à quel point la notion de prototype est importante à ce stade. C'est pourquoi nous proposons un système de requête non pas par l'exemple, mais via la construction d'un *modèle* (voir section 3.2.1.1). Ceci permet, en outre, un gain de souplesse dans le processus de requête.

La notion de similarité ne peut se réduire à une distance

Enfin, nous avons vu que la notion même de similarité est rapidement escamotée dans la plupart des travaux pour installer une indexation basée sur une distance mathématique. Or, cette dernière s'est révélée inadaptée pour modéliser un certain nombre de cas, concernant le jugement humain de similarité. Conscients de cette limitation, nous proposerons donc dans cette thèse une mesure de similarité particulière, plus en accord avec les résultats de Tversky. Dans l'optique de la recherche d'objets dans les bases d'images, cette mesure de similarité intègrera notamment une forte composante structurelle (voir section 3.4).

La figure 1.29 synthétise la chaîne de traitements proposée dans cette thèse. Il convient de bien noter que nous avons implémenté un prototype sur cette base, qui permet de prendre en charge la totalité du processus d'indexation et de requête à une base d'images.

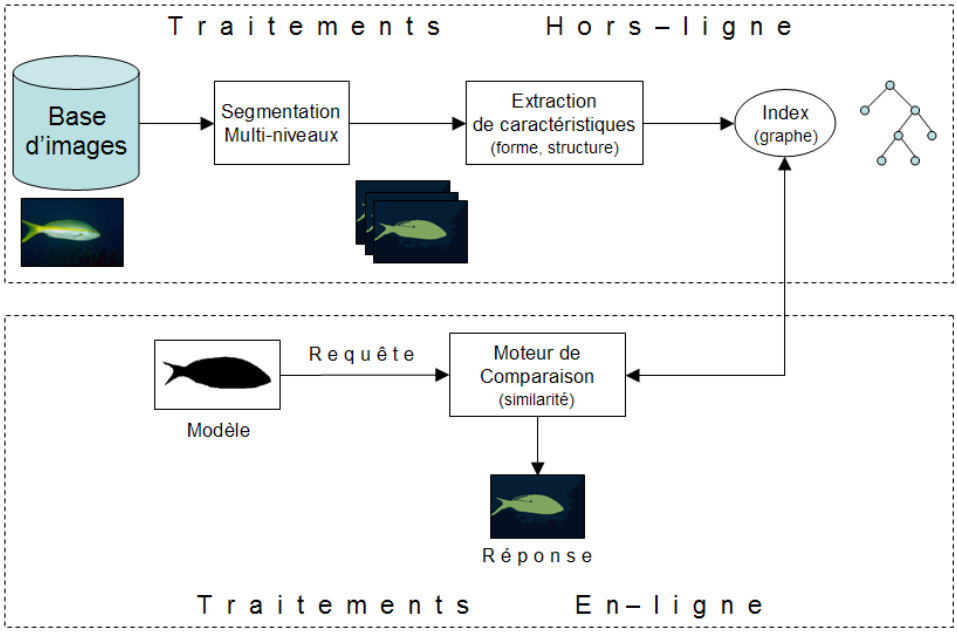


FIG. 1.29 – Chaine de traitements pour l'indexation, proposée dans cette thèse.

2

De la segmentation au groupement perceptuel

Sommaire

2.1 Introduction

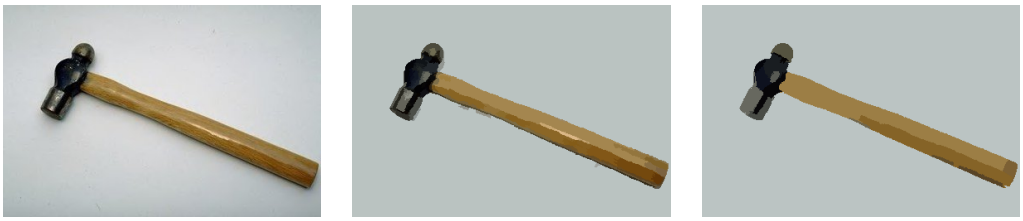
Nous avons vu dans la partie précédente que la segmentation est au centre d'une activité intense de recherche. Le problème n'est pas nouveau en indexation. Au contraire, il renvoie à des questions qui s'étaient déjà posées dans le domaine de la vision par ordinateur. Cette dernière cherche à extraire des images une description symbolique. Sa finalité n'est donc pas strictement identique à la problématique de l'indexation. Néanmoins, il est évident que les mécanismes mis en jeu peuvent être, pour une large part, communs.

Une étape cruciale en vision concerne ce qu'il est convenu d'appeler le groupement perceptuel (Lowe, 1985), en référence à la capacité que possède le système visuel humain d'imposer structure et régularité à des stimuli variés. Le groupement perceptuel vise donc à extraire des structures saillantes de l'image, qui pourront ensuite être manipulées par des tâches interprétatives de plus haut niveau, en vue de reconnaître des objets du monde 3D. Les avantages de manipuler de telles structures au lieu des simples pixels sont nombreux : tout d'abord, la complexité des traitements futurs se voit considérablement réduite (par exemple s'il s'agit de comparer les structures extraites à des bases de modèles d'objets, au lieu de tester toutes les partitions possibles de pixels). Ensuite, le fait de manipuler des primitives de type région ou contour permet de disposer d'outils plus proches du niveau objet, comme la forme ou les relations spatiales. Or, nous avons déjà vu dans la partie 1 que ces outils peuvent s'avérer particulièrement expressifs.

2.1.1 Groupement perceptuel et segmentation

La forme la plus immédiate de groupement perceptuel est naturellement la segmentation faible, qui consiste à grouper les pixels en régions, selon des critères bas-niveau comme la couleur ou la texture (Carson et al., 2002; Wang et al., 2001; Comaniciu et Meer, 1997). On aboutit ainsi à une partition de l'image en régions, chacune d'entre elles étant homogène en termes de couleur ou de texture, selon le critère utilisé. Toutefois, ces approches sont très limitées car elles ne permettent pas d'extraire des régions composées de plusieurs parties avec pour chacune des descripteurs bas-niveaux différents. Or, un grand nombre d'objets réels sont de ce type.

De plus, la segmentation est, par nature, très sensible aux différents artefacts issus des conditions d'éclairage ou des occlusions par exemple. A titre d'illustration, la figure 2.1 présente deux résultats de segmentation par mean-shift (Comaniciu et Meer, 1997) pour une même image originale (a). Les différents résultats ont été obtenus en faisant varier les paramètres de granularité de la segmentation. On remarque que les différentes parties de l'objet ont été découpées en plusieurs régions, du fait des reflets et aux autres variations de l'éclairage.



(a) (b) (c)

FIG. 2.1 – Exemple de différentes segmentations sur une image originale (a).

2.1.2 Vers d'autres propriétés de groupement perceptuel : l'école Gestalt

Pour réduire ces problèmes d'artefacts, des travaux ont tenté d'appliquer des propriétés supplémentaires lors du groupement perceptuel. Par exemple, il pourrait être intéressant de grouper deux régions dont les frontières présentent une certaine continuité. En effet, cela pourrait signifier que ces deux régions sont issues du même objet. Ainsi, sur l'image de la figure 2.1 (c), on souhaiterait intuitivement regrouper les deux grandes régions du manche, du fait de la continuité qui existe entre leur contours.

Or, au cours du XXème siècle, des travaux de psychologues de l'école Gestalt, tels KOFFKA (1935) ou plus tard KANIZSA (1979), avaient déjà mis en évidence un certain nombre de propriétés de groupement de stimuli, à l'oeuvre durant la perception de forme par l'humain.

Parmi celles-ci, on peut citer, par exemple, la proximité, la continuité, le parallélisme, ou la fermeture. La figure 2.2 illustre certaines de ces lois. Par exemple, en (a), nous avons naturellement tendance à considérer la figure comme étant composée de cinq colonnes principales, sous l'effet de regroupement des cercles adjacents en une seule colonne centrale. En (b), bien que tous les cercles soient équidistants, nous percevons la figure comme étant composée d'une colonne centrale de cercles noirs, au milieu d'un fond uniforme. La propriété de fermeture (c) possède, quant à elle, un champ d'application plus vaste que ce que son nom seul pourrait laisser présager : elle stipule que la perception cherche à former des objets simples, fermés et compacts. Ainsi, les deux formes imbriquées de la figure (c) auront tendance à être regroupées en une seule, plus compacte. L'exemple (d) illustre la propriété de continuité, qui cherche à regrouper deux entités dont les contours présentent un raccord "lisse". Ainsi, les deux formes en (d) seront plutôt perçues comme une seule entité que comme deux figures adjacentes, aux bords saillants. La figure (e) illustre un mécanismes de regroupements de deux formes en une seule, sous l'action d'une propriété de parallélisme entre les contours.

Notons enfin l'exemple (f), plus complexe, qui rend compte d'un phénomène d'illusion : bien qu'aucun contour ne dessine explicitement de triangle, celui-ci ap-

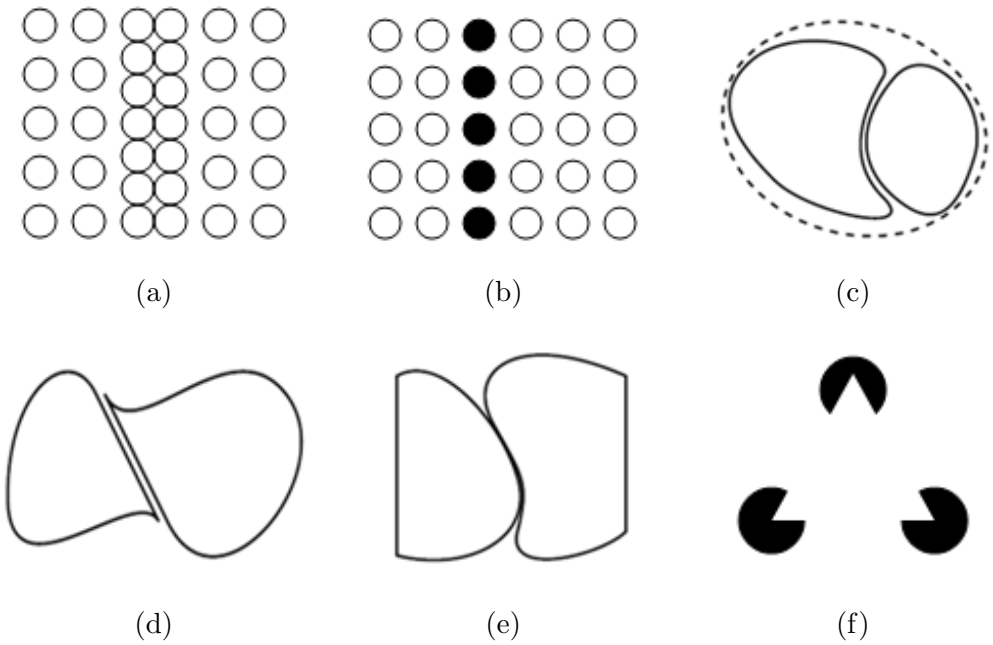


FIG. 2.2 – Exemple de différentes propriétés Gestalt à l'oeuvre : proximité (a), similarité (b), fermeture (c), continuité (d), parallélisme (e) et triangle de Kanisza (f).

parait en recouvrement partiel des cercles noirs.

Par la suite, de nombreux travaux issus du domaine de la vision par ordinateur ont proposé des implémentations à ces propriétés (LOWE, 1985; SARKAR et BOYER, 1993b). En effet, l'utilisation de propriétés géométriques lors du groupement perceptuel, en plus de propriétés locales de couleur ou de texture, permet d'améliorer la qualité de la segmentation et de s'affranchir de plusieurs problèmes d'artefacts.
La plupart des travaux existants sur le groupement perceptuel utilisent des approches basées sur une extraction préalable des contours de l'image. De plus, elles utilisent une seule propriété à la fois pour activer un groupement.

Nous avons proposé durant cette thèse une nouvelle approche au goupement perceptuel, basée sur des primitives régions. Ces dernières nous permettent d'utiliser ainsi des informations aussi bien région que contour pour caractériser les groupements. En outre, nous demandons l'activation simultanée de plusieurs propriétés Gestalt pour déclencher un groupement. Ceci représente un certain gain de robustesse, en empêchant une propriété seule d'activer un groupement.

Le plan de cette partie sera le suivant : tout d'abord, nous allons présenter un état de l'art des approches de groupement perceptuel (section 2.2). Ensuite, nous présenterons notre modélisation du groupement perceptuel, qui comportera

deux volets : l'interaction entre les propriétés par la théorie de l'évidence (section 2.3.2) et la description des propriétés en tant que telles (section 2.3.3). Enfin, nous présenterons différents résultats (section 2.4).

2.2 Etat de l'art en groupement perceptuel

Les mécanismes de groupement perceptuel peuvent être divisés en deux types de traitement, toujours en référence au système visuel humain : les processus pré-attentifs d'une part, et attentifs d'autre part.

2.2.1 Groupement perceptuel attentifs

Les processus attentifs (ou *top-down* ou encore descendants) s'appuient sur des connaissances externes afin d'effectuer des tâches avec un but précis (par exemple : reconnaître une forme particulière, dans une zone fixée de l'image). Ces connaissances peuvent être de nature et de forme variées. Par exemple, des réseaux bayesiens ont été utilisés par Sarkar et Boyer (1993a) afin d'inférer la présence ou l'absence de figures géométriques (rectangle, carré...) à partir de différents indices (coins, courbes, segments, segments parallèles...).
Les connaissances externes peuvent également prendre la forme de modèle d'objets. Ainsi, Sclaroff et Liu (2001) groupent des régions en utilisant des modèles statistiques de formes. Ces derniers leur permettent de déduire des probabilités a posteriori de déformations globales pour chaque classe d'objets (une classe étant par exemple, le type d'objet *feuille*, ou *poisson*, etc...). Dans la même optique, Forsyth et Fleck (1999) groupent des régions de couleur chair suivant des structures pré-établies, afin de reconnaître des corps nus d'humains.
D'une manière générale, il est difficile de généraliser de tels systèmes, car ils reposent souvent sur un certain nombre de traitements ad-hoc.

Les processus pré-attentifs (ou *bottom-up* ou encore ascendants) manipulent quant à eux uniquement les données bas-niveaux (signal) issues de l'image, sans connaissances externes. Il est évident qu'un système efficace de vision se doit d'intégrer les deux types de processus. Dans cette partie, nous nous focalisons sur les traitements pré-attentifs. Nous verrons par la suite dans le chapitre 3 comment il est possible de mettre en oeuvre des processus attentifs, afin de vérifier et d'interpréter les résultats des traitements pré-attentifs.

2.2.2 Groupement perceptuel pré-attentifs

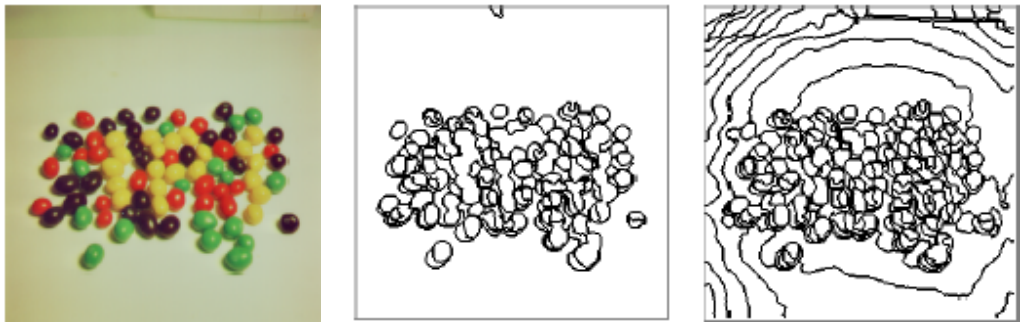
2.2.2.1 Le type de primitives : régions ou contours

Sarkar et Boyer (1993b) présentent un important état de l'art du groupement

perceptuel. Ils constatent que la majorité des approches utilisent les propriétés Gestalt de proximité, similarité, fermeture, continuité et parallélisme, sur des primitives de type contour. Or, utiliser de manière exclusive ce type de primitives apparaît comme une limitation. En effet, chaque primitive se montrera adaptée à un certain nombre de propriétés, mais jamais à l'ensemble d'entre elles. Ainsi, les régions permettent, par exemple, de définir une notion de similarité du point de vue de descripteurs couleur ou texture, ce que ne peut produire une primitive contour. A l'inverse, les propriétés de continuité ou de parallélisme sont plus facilement implémentables sur des primitives contours, puisqu'elles sont relatives à l'orientations de segments.

Si autant de travaux se basent sur les contours, c'est en partie parce qu'il a été admis très tôt qu'ils constituent une primitive essentielle à la vision humaine (Marr, 1982), par le biais d'une extraction différentielle de type *zero-crossing*. Néanmoins, les primitives régions grossières, ou *blobs* ont depuis été reconnues comme également fondamentales dans les traitements pré-attentifs (Oliva, 2005).

Les primitives régions sont également souvent négligées, car elles induisent lors de leurs extractions un certain nombre d'artefacts liés à la confusion objet / fond. En effet, lors d'une segmentation région en couleur par exemple, il est courant de voir un objet fusionné avec son fond si leurs couleurs sont similaires. A l'inverse, une extraction contour, sensible aux contrastes locaux, sera plus robuste dans ce type de traitement et permettra de mieux séparer un objet de son fond. La figure 2.3 présente un exemple de segmentation couleur en région par un algorithme classique de type *mean-shift* (Comaniciu et Meer, 1997). Notons que dans un souci de précision, les images présentent les contours des régions extraites.



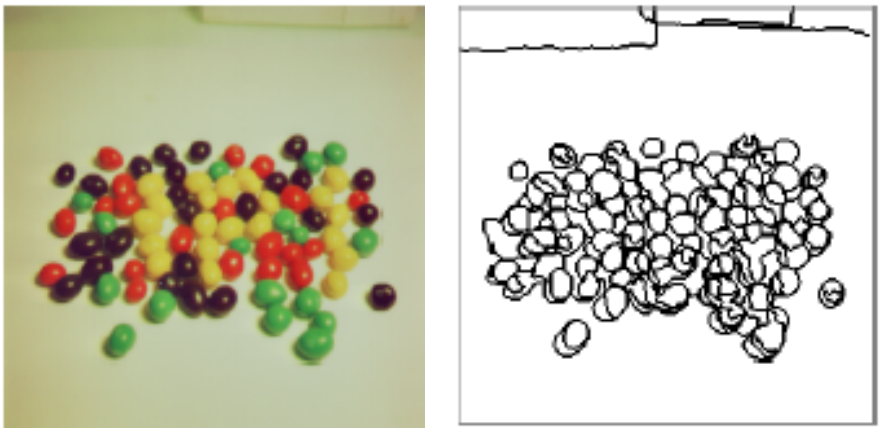
(a) (b) (c)

Fig. 2.3 – Exemple de segmentation couleur en régions par *mean-shift* (b) et sur-segmentation (c) sur une image originale (a).

On remarque que l'algorithme ne parvient pas à séparer les différentes billes jaunes à gauche de l'image : elles sont toutes fusionnées dans une seule région (figure 2.3-b). En augmentant la granularité de la segmentation (figure 2.3-c), il est possible de séparer les billes mais au prix d'un bruit général très important.

Une amélioration consiste à introduire une information supplémentaire sur les contours. En effet, une segmentation classique en régions s'appuie sur une quantification d'un descripteur bas-niveau donné (couleur, texture) en chaque pixel de l'image. Ensuite, le processus consiste à regrouper les pixels proches, possédant des valeurs de descripteur similaires. Christoudias et al. (2002) proposent, lors de ce regroupement, d'intégrer une information supplémentaire, relative au contours extraits de l'image. Ces derniers correspondent aux zones de fort gradient de l'image. Plus précisément, l'idée consiste à ne pas regrouper des pixels ayant des descripteurs bas-niveaux proches, mais séparés par un contour marqué. Ce mécanisme permet alors d'être moins sensible aux artefacts classiques des méthodes régions, résultant de la confusion objet/fond.

La figure 2.4 présente un exemple de résultat sur la même image que précédement. Les billes sont séparées et le bruit induit reste très faible comparé à la sursegmentation (c) de la figure 2.3.



(a) (b)

Fig. 2.4 – Exemple de segmentation couleur en régions en intégrant une information contour (b) sur une image originale (a).

Dans la suite, nous nous appuierons sur une segmentation couleur préalable, par la méthode du *mean-shift* (Comaniciu et Meer, 1997), en intégrant une information contour. Cela nous permettra de manipuler des primitives de type région, indispensables pour quantifier certaines propriétés. Néanmoins, nous éviterons les artefacts classiques de la segmentation région, selon la méthode préconisée par Christoudias et al. (2002). En outre, cela nous permettra d'avoir recours directement aux contours des régions pour quantifier certaines propriétés.

2.2.2.2 Le principe de non-accidentalité

Lowe (1985) est le premier à avoir utilisé des propriétés Gestalt pour effectuer

des groupements perceptuels. Une de ses contributions importante est l'introduction du principe de non-accidentalité, afin de calculer la saillance d'un groupement : celle-ci est définie comme inversement proportionelle à la probabilité d'apparition. Ceci signifie que nous jugeons saillante, une structure en contraste avec son environnement, et non pas en tant que telle. Ceci est corroboré par un certain nombre d'études sur la perception humaine. La figure 2.5 présente quelques exemples bien connus, d'après Hansen (2002). Ainsi, dans la figure (a), les deux petits carrés gris clairs, placés à l'intérieur d'un carré plus large, ont la même luminance. Pourtant, ils apparaissant teintés différement, à cause de la différence de luminance des carrés plus larges qui les englobent. Leur saillance respective varie donc en fonction de leur environnement. De même, en (b), les deux bords droite et gauche du carré ont même luminance, mais apparaissent différents à cause de la zone de fort contraste au centre.

(a) (b)

Fig. 2.5 – Illustration du mécanisme de saillance par contraste, d'après Hansen (2002).

Toutefois, l'utilisation de ce principe par Lowe (1985) reste peu formalisé. On peut également citer Desolneux et Morel (2003) qui ont formalisé un raisonnement analogue, à partir du principe d'Helmholtz, pour calculer la saillance d'un groupement.

Plus tard, Oliva et al. (2003) ont directement formalisé la saillance S d'un site x de l'image, décrit par un vecteur de caractéristiques $v = \{v(x)_{k,0\leq x\leq N}\}$ de dimension N, par l'inverse de la probabilité d'apparition du vecteur :

$$S(x) = \frac{1}{p(v)} \tag{2.1}$$

La probabilité $p(v)$ est approximée par une loi gaussienne :

$$p(v) = \frac{1}{(2\pi)^{N/2}\,|X|^{1/2}} e^{-\frac{1}{2}(v-\mu)^T X^{-1}(v-\mu)} \tag{2.2}$$

Dans le cas le plus simple, le vecteur de caractéristiques v associé au pixel peut être son niveau de gris. La saillance d'un site est alors inversement proportionnelle à la probabilité d'apparition de son niveau de gris. Autrement dit, plus un pixel présente un niveau de gris "rare", plus ce dernier sera saillant.

Nous verrons dans la suite comment nous avons proposé une extension de cette notion de saillance d'un site.

2.2.2.3 Les hiérarchies de groupement

Une fois dressé leur état de l'art, Sarkar et Boyer (1994) proposent une hiérarchie de groupement, basée sur une segmentation préalable en contours. Ainsi, des structures (*tokens*) de complexité croissante sont regroupées itérativement, selon une hiérarchie préalablement fixée. Des structures à base de graphes, dont les noeuds sont les différentes structures, sont utilisées. Des opérations sur ces graphes permettent alors d'extraire d'autre tokens : par exemple le plus court chemin entre deux noeuds dans un graphe dont les noeuds sont des segments adjacents, permet d'extraire des formes fermées. Dans la même idée, les composantes connexes sur des graphes de segments proches extraient des jonctions. La figure 2.6 présente la liste des opérations effectuées sur les différents graphes afin d'obtenir des structures variées.

Des approches similaires ont été proposées (Ackermann et al., 1997; Kang et Walker, 1994; Mohan et Nevatia, 1992) avec différents contrôles globaux sur les groupements, comme par exemple des champs de Markov (Ackermann et al., 1997) ou des logiques floues (Kang et Walker, 1994).

Toutefois, ces travaux extraient toujours des structures prédéfinies, de manière statique, ce qui est un facteur limitant. De plus, ils utilisent une seule propriété Gestalt à la fois pour caractériser un groupement, alors que plusieurs propriétés sont souvent à l'oeuvre en coopération, afin de réaliser un groupement (Desolneux et Morel, 2003). Prendre cette information en considération peut aboutir à un gain important en robustesse, puisque cela empêche une loi d'activer un groupement à elle seule. Nous modélisons une telle coopération dans notre architecture.

2.2.2.4 Les interactions entre les propriétés Gestalt

Afin de prendre en compte des interactions plus complexes entre les propriétés, et sans aucune hiérarchie statique de groupement, Murino et al. (1996) modélisent un graphe avec pour noeuds les contours extraits préalablement d'une image et pour arcs les hypothèses de groupement entre ces contours. La figure 2.7 présente un exemple de tel graphe entre sept contours extraits. On constate qu'entre deux contours (noeuds) donnés, une seule hypothèse (arc) peut être instantiée. Par exemple, l'hypothèse de colinearité a été instantiée entre les noeuds 1 et 2.

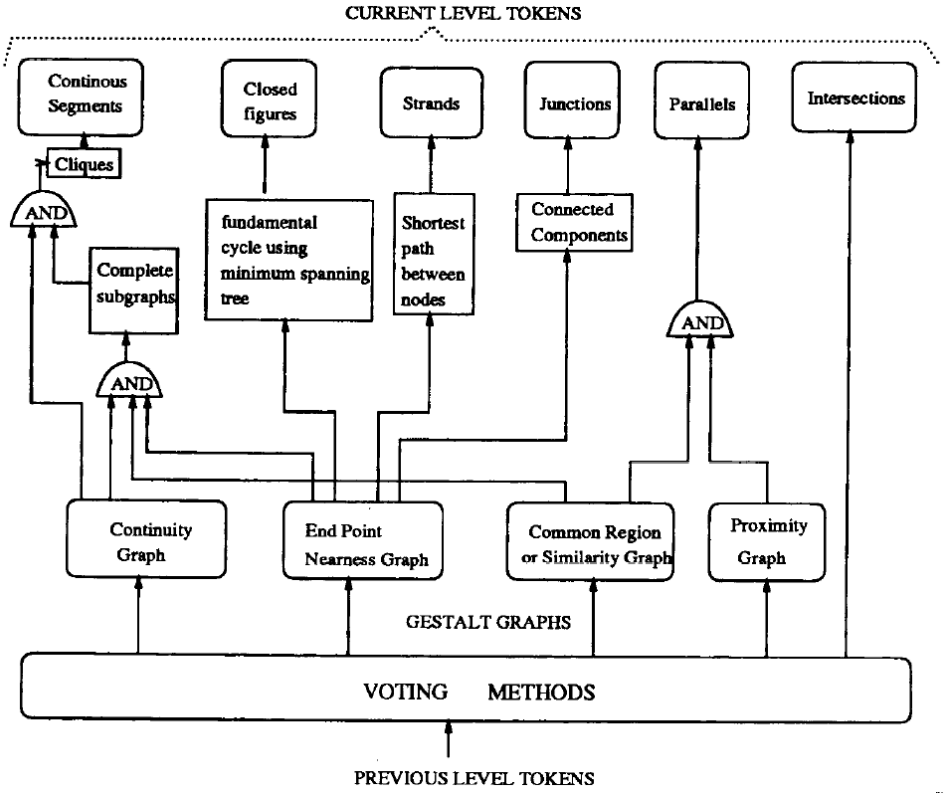


FIG. 2.6 – Hiérarchie pré-attentive de groupement de SARKAR et BOYER (1994)

Chaque arc est valué par une probabilité de groupement selon une propriété Gestalt donnée. Une modélisation par champ de Markov permet ensuite de minimiser l'énergie du système et donc de trouver son état le plus stable. Ceci va induire une réduction du graphe et donc le regroupement de différents noeuds (segments). Mais encore une fois, une seule propriété Gestalt est utilisée à la fois pour chaque groupement.

IDRISSI ET AL. (2004) utilisent un algorithme glouton pour réduire un graphe analogue basé région, dans lequel chaque hypothèse de groupement est caractérisée par les propriétés de proximité, similarité et fermeture. Néanmoins, ils utilisent une somme pondérée pour combiner les influences des différentes propriétés et ne modélisent donc pas finement l'interaction. Une approche similaire est proposée par LUO et GUO (2003).

En fait, les approches bayesiennes exposées précédemment ne sont pas particulièrement adaptées pour modéliser finement l'interaction de différentes propriétés Gestalt au sein d'une hypothèse de groupement. En effet, il arrive fréquemment qu'au moins une partie de ces propriétés se contredisent localement. Il est alors impossible de conclure sur l'hypothèse. Au contraire, la théorie de l'évidence de Dempster-

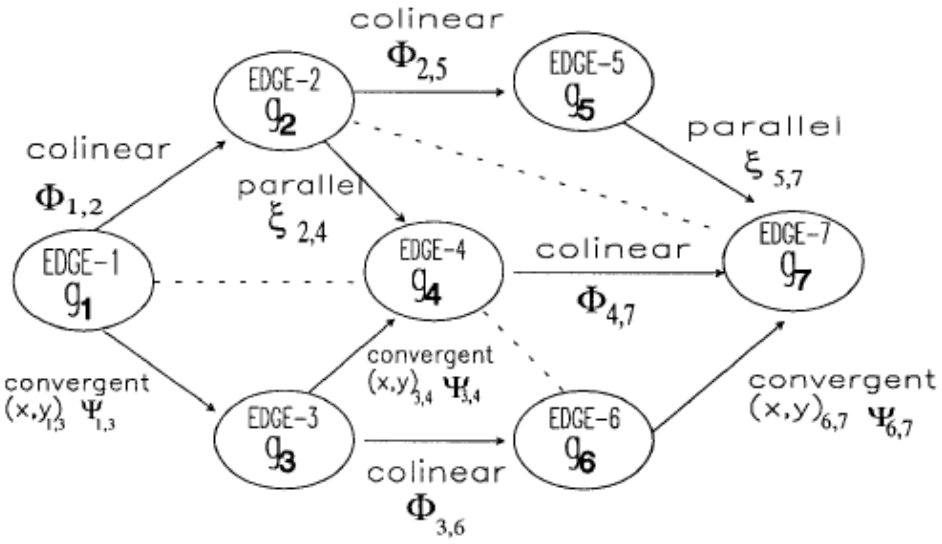


FIG. 2.7 – Graphe de groupement entre segments (MURINO ET AL., 1996)

Shafer (SHAFER, 1976) a été spécifiquement conçue pour un tel cas. Cette dernière modélise, sur une hypothèse, la notion de *croyance* plutôt que de probabilité. La différence fondamentale entre ces deux concepts réside dans le fait qu'il est possible d'allouer une portion de croyance x à une hypothèse de type groupement sans pour autant allouer le reste $(1 - x)$ à sa négation. Ainsi, différentes propriétés sur une hypothèse peuvent être combinées sans pour autant se contredire. Il est en outre possible de modéliser la notion d'incertitude résiduelle sur l'hypothèse. VASSEUR ET AL. (1999) ont proposé d'utiliser la théorie de l'évidence pour grouper des primitives contour dans une image. Toutefois, ces travaux se limitent à une vue bayesienne de la théorie de l'évidence, et conduisent à des propriétés en contradiction les unes avec les autres. Comme nous le verrons par la suite, notre architecture modélise une réelle coopération des propriétés entre elles, et empêche un conflit de bloquer le processus de groupement.

2.2.3 Les approches pré-attentives psycho-visuelles

Comme exposé dans la partie 1, de nombreux travaux se sont attachés à modéliser l'attention pré-attentive du système visuel humain. Ces derniers s'appuient sur des résultats de travaux issus des neurosciences, de la biologie et des sciences cognitives. Ils cherchent à recréer, par le biais d'opérations à base de filtres, les processus de la perception visuelle qui ont lieu depuis la rétine jusqu'au cortex inférieur. Dans ce cadre, le modèle de ITTI ET AL. (1998) fait référence. Nous le présentons rapidement maintenant.

2.2.3.1 Le modèle de Itti et al. (1998)

Comme illustré dans la figure 2.8, ce modèle crée trois canaux d'entrée à partir d'une image couleur r,v,b :

- un canal d'intensité $I = \frac{r+v+b}{3}$
- un canal couleur composé de quatre composantes, liés à la théorie des couleurs antagonistes : rouge, vert, bleu, jaune
- un canal de composantes orientées, issues du canal intensité d'une pyramide de Gabor orientée (σ, θ) avec σ le niveau de la pyramide et $\theta \in \{0, 45, 90, 135\}$ l'orientation du filtre

Une décomposition sur 9 niveaux est effectuée sur chacune des composantes, via des pyramides gaussiennes. Cette décomposition permet de créer une carte de caractéristiques pour chaque composante. La combinaison de ces différentes cartes fournit alors une carte de saillance.

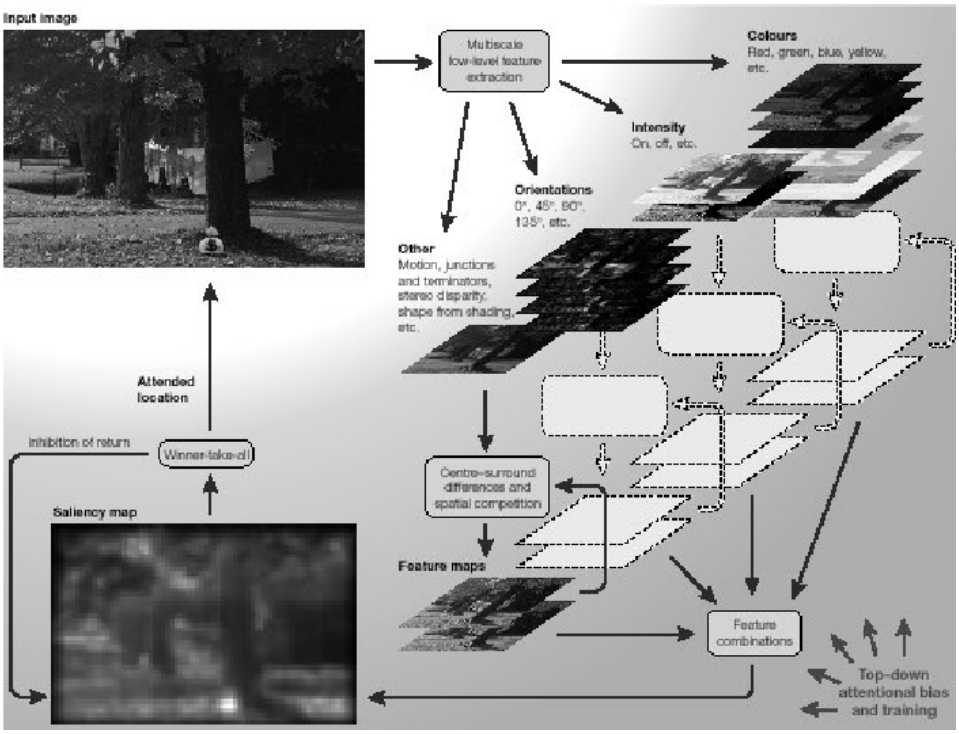
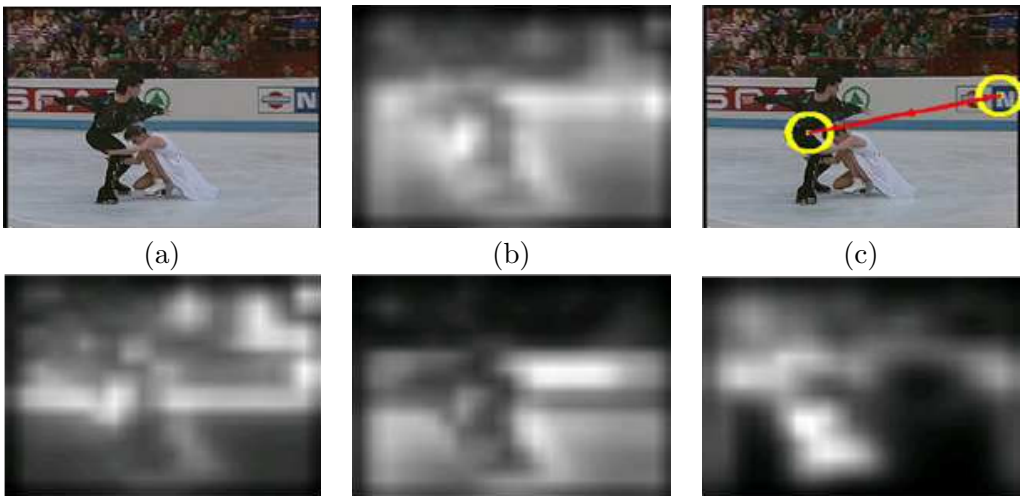


Fig. 2.8 – Modèle d'attention visuel pré-attentive de Itti et al. (1998)

La figure 2.9 donne un exemple de résutat du modèle Itti et al. (1998)



(a) (b) (c)

(d) (e) (f)

FIG. 2.9 – Exemple de résultat du modèle ITTI ET AL. (1998) à partir d'une image (a) : saillance globale (b), deux points d'attention principaux (c) et cartes de saillances couleur (d), intensité (e), orientation (f). Reproduit de LE MEUR ET AL. (2004).

2.2.3.2 Les autres modèles

D'autres modèles d'attention visuelle ont été proposés. Citons par exemple celui de MILANESE ET AL. (1992), qui se focalise sur deux limitations du modèle précédent. La première concerne la fusion des différentes cartes de saillance en une carte unique. On retrouve la problématique plus générale de fusion de données sur une hypothèse, évoquée précédemment. La deuxième limitation concerne le résultat même du modèle ITTI ET AL. (1998). En effet, la carte de saillance obtenue n'est pas binaire. Il faut donc seuiller cette dernière pour obtenir une liste de points saillants ou de régions saillantes. Milanese propose donc un critère de fusion des cartes de saillance, basé sur une moyenne, ainsi qu'un certain nombre de pré-traitements. Parmi ceux-ci peut être cité l'utilisation d'un filtre passe-bas isotrope gaussien afin de privilégier l'apparition de forme compacte. On retrouve le critère de fermeture.

2.2.3.3 Positionnement par rapport aux approches psycho-visuelles

Les modèles d'attention visuelle ont permis d'obtenir des résultats de qualité, notamment en se confrontant à des expériences occulométriques (enregistrement des points de fixation d'un opérateur humain). Néanmoins, ce type de traitements aboutit à l'extraction de régions grossières ou blobs. Ces dernières ne sont pas suffisamment précises pour permettre ensuite de les caractériser par des descripteurs d'objets comme la forme. Or, c'est ce que nous recherchons dans une perspective

d'indexation.

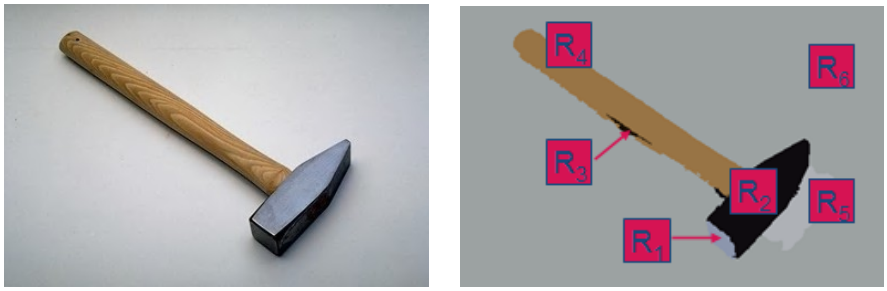
C'est pourquoi nous n'avons pas utilisé directement des approches psycho-visuelles. Néanmoins, dans l'optique d'extraire des zones de l'image pertinentes pour un utilisateur humain, nous avons considéré les propriétés de groupement du système visuel humain au niveau macroscopique (Gestalt). Puis, nous en avons proposé une modélisation fonctionnelle, afin de pouvoir les appliquer à des images numériques. En résumé, nous n'avons pas cherché à modéliser le système visuel humain, mais plutôt à extraire d'images des structures visuelles pertinentes pour l'humain.

2.3 Modélisation du groupement perceptuel

Nous proposons dans cette thèse une architecture coopérative de groupement perceptuel que nous allons maintenant présenter. Nous utilisons une image sur-segmentée en régions comme point de départ, et nous imposons à plusieurs propriétés Gestalt d'être activées afin de déclencher un groupement. Ceci permet un certain gain en robustesse car il empêche une propriété de déclencher à elle seule un groupement.

2.3.1 Processus général

Les hypothèses de groupement sont générées à partir d'un graphe d'adjacence (RAG), dans lequel les noeuds représentent les régions issues de la sur-segmentation préalable et les arcs, une adjacence entre les régions correspondantes. Ainsi, les différents arcs représentent les hypothèses de groupement. La figure 2.10 présente un exemple simplifié de sur-segmentation préalable (b) à partir d'une image originale (a).



(a) Image originale (b) Image segmentée

FIG. 2.10 – Exemple simplifié de sur-segmentation préalable.

Le graphe d'adjacence correspondant à la sur-segmentation (b) de la figure 2.10 est représenté à la figure 2.11.

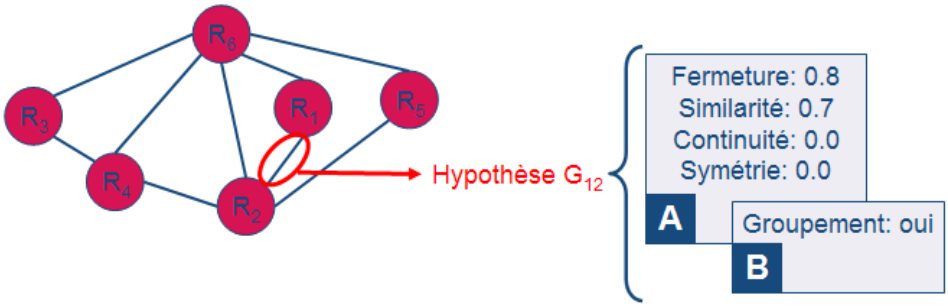


FIG. 2.11 – Exemple simplifié de graphe d'adjacence correspondant à la sur-segmentation de la figure 2.10.

Pour chaque hypothèse de groupement, chaque propriété Gestalt induit une croyance partielle en sa réalisation (étape A de la figure 2.11). Puis, les croyances partielles sont combinées par la méthode de l'évidence en une croyance globale sur l'hypothèse de groupement (étape B). Le graphe est ensuite réduit itérativement, à partir de ces croyances globales. La figure 2.12 illustre le processus de groupement perceptuel.

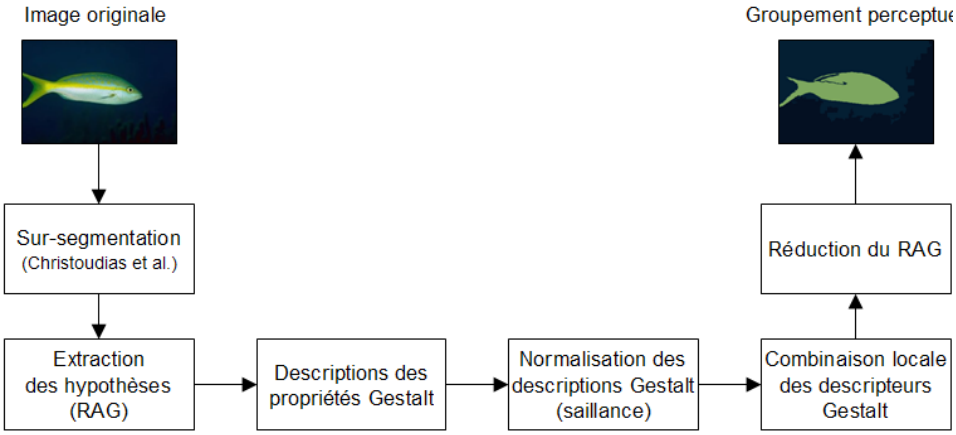


FIG. 2.12 – Processus du groupement perceptuel.

Nous détaillons dans les parties suivantes le modèle d'interaction utilisé pour combiner l'influence des propriétés Gestalt (section 2.3.2), puis la mesure proprement dite des différentes propriétés, c'est-à-dire l'extraction de descripteurs correspondants (section 2.3.3) et la normalisation de ces derniers par saillance (section 2.3.4).

2.3.2 Modèle d'interaction des propriétés Gestalt

La principale contribution de ce travail réside dans la combinaison des influences des différentes propriétés Gestalt sur chaque hypothèse, afin d'en déduire pour chacune une valeur quantitative de confiance. Nous l'avons vu, une solution immédiate consiste souvent à introduire une somme pondérée. Néanmoins, cette méthode, outre son caractère fortement heuristique, n'est pas robuste à l'erreur locale. Ainsi, s'il advient qu'une seule propriété Gestalt ne réponde pas sur une hypothèse, la confiance globale en celle-ci est fortement pénalisée. C'est précisément dans cette dynamique qu'a été introduite la théorie de l'évidence par Shafer (1976). En effet, elle s'attache à combiner différents points de vue sur une hypothèse. Ici, les hypothèses sont les groupements perceptuels possibles entre deux régions adjacentes R_i et R_j. Chaque propriété Gestalt peut être considérée comme un point de vue différent sur chacune des hypothèses. La théorie nous permet alors de dériver un point de vue résultant, qui prend en compte tous les autres.

Nous allons maintenant présenter rapidement le formalisme de Dempster-Shafer. Pour une vision plus détaillée, le lecteur est invité à se reporter aux annexes.

2.3.2.1 Notations et définitions de la théorie de l'évidence

Cette théorie, probabiliste, modélise des croyances sur des jeux d'hypothèses, plutôt que des probabilités (Shafer, 1976).

Cadre de discernement

Soit Θ un ensemble d'hypothèses mutuellement exclusives, $\{H_1, H_2, \ldots, H_n\}$, appelé cadre de discernement. L'ensemble de toutes les parties de Θ est noté 2^Θ.

Jeu de masses m

On appelle jeu de masses une fonction $m : 2^\Theta \longrightarrow [0\,;1]$ qui satisfait :

$$m(\varnothing) = 0 \tag{2.3}$$

$$\sum_{A \subset \Theta} m(A) = 1 \tag{2.4}$$

Pour chaque partie A de Θ, $m(A)$ représente la croyance que quelqu'un engage exactement en A. De manière simplifiée, m peut être vue comme une fonction de probabilité "améliorée".

Construire un jeu de masses m, consiste à répartir toute la croyance disponible sur le jeu d'hypothèses H_i de Θ ou, plus généralement, sur toute sous-partie A de Θ. Par exemple, étant donné $\Theta = \left\{H_1, \overline{H_1}\right\}$, je peux poser $m(H_1) = x$, qui signifie que je crois H_1 au degré x. L'originalité du formalisme réside dans le fait qu'à ce stade,

je ne suis pas obligé d'attribuer toute la croyance restante $1 - x$ sur l'hypothèse contraire $\overline{H_1}$. Au contraire, je peux dire que je ne crois pas du tout en l'hypothèse $\overline{H_1}$,c'est à dire : $m(\overline{H_1}) = 0$. Par souci de normalisation (equation 2.4), je pose alors $m(\Theta) = 1 - x$. Cette croyance résiduelle (je crois à l'espace des possibles au degré $1 - x$) est appelée incertitude.

Comparaison avec les probabilités

A titre de comparaison, une fonction de probabilité classique p aurait elle aussi pu définir $p(H_1) = x$. Mais il s'en serait alors automatiquement suivi que $p(\overline{H_1}) = 1 - x$.

La théorie de l'évidence autorise donc une modélisation plus fine des croyances sur un jeu d'hypothèses, que ne le font les probabilités classiques. La figure 2.13 montre deux exemples de jeux de masses, m_1 et m_2, sur le cadre de discernement $\Theta = \{H_1, H_2\}$.

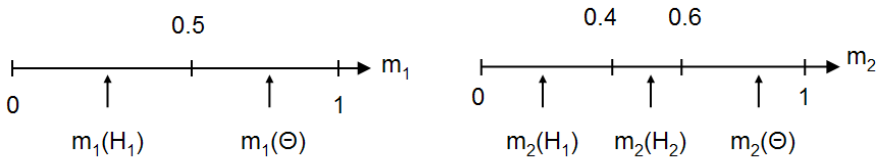


FIG. 2.13 – Exemple de deux jeux de masses sur le cadre de discernement $\Theta = \{H_1, H_2\}$.

Lorsqu'un jeu de masses m engage une croyance en une partie A de Θ, on a $m(A) > 0$ et A est appelé un élément focal de m.

D'une manière théorique, la théorie de l'évidence distingue la croyance *exacte* que l'on porte en un évènement (avec un jeu de masses) de la croyance *totale* (avec d'autres types de fonctions). Nous n'insistons pas sur cette distinction ici, dans un souci de simplification. Une plus ample description de la théorie est donnée en annexes.

2.3.2.2 Combinaison des jeux de masses : la règle de Dempster

La théorie de l'évidence nous fournit alors la règle de Dempster, qui permet de combiner différents jeux de masses sur un même cadre de discernement, afin d'en déduire un nouveau jeu, prenant en compte l'influence de tous les autres.

Exemple graphique

Considérons dans un premier temps, mais sans limitation aucune, le cas à deux jeux de masses de la figure 2.13. La combinaison peut être illustrée par la figure 2.14.

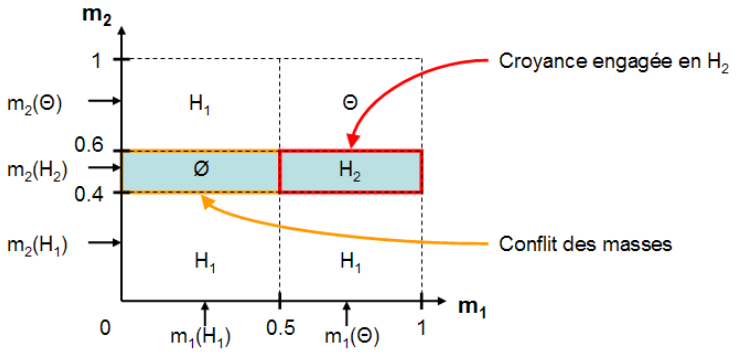


FIG. 2.14 – Combinaison des deux jeux de masses de la figure 2.13. La croyance engagée en chaque cas est représentée par l'aire de la région associée.

Elle donne lieu à six cas possibles. La croyance résultante affectée à chacun des cas est représentée graphiquement par l'aire correspondante. Par exemple, il est possible de calculer la croyance résultante en H_2 :

$$m(H_2) = m_2(H_2)m_1(\Theta) \tag{2.5}$$

Formalisation de la règle de Demspter

Ceci peut être formalisé ainsi : soient m_1 and m_2 deux jeux de masses sur le même cadre de discernement Θ. Soient A_i les éléments focaux de m_1 et B_j ceux de m_2. On appelle somme orthogonale de m_1 et m_2, le jeu de masses m défini sur toutes les parties C de Θ par :

$$m(\oslash) = 0 \qquad m(C) = \frac{\sum\limits_{A_i \cap B_j = C} m_1(A_i)m_2(B_j)}{1 - \sum\limits_{A_i \cap B_j = \oslash} m_1(A_i)m_2(B_j)} \tag{2.6}$$

La somme orthogonale de m_1 et m_2 représente la croyance combinée que quelqu'un a sur Θ, étant donnés m_1 et m_2.

Notion de conflit

Puisque le cadre de discernement est composé d'hypothèses mutuellement exclusives, il peut arriver que la règle de combinaison donne des sous-cas correspondant à une intersection vide ($\oslash$). On dit qu'il y a conflit entre les jeux de masses m_1 et m_2. On peut quantifier ce conflit avec le nombre k défini par :

$$k = \sum_{A_i \cap B_j = \oslash} m_1(A_i) m_2(B_j) \tag{2.7}$$

Lorsque le nombre k augmente, le conflit est important et le résultat de l'équation (2.6) risque de ne plus être représentatif. Lorsque $k = 1$, le conflit est total et la somme orthogonale n'existe pas.

Exemple

Ainsi, pour les deux jeux de masses de la figure 2.13, on a, après combinaison :

$$k = 0.6 \cdot 0.2 = 0.12 \tag{2.8}$$

$$m(H_1) = \frac{0.5 \cdot 0.4 + 0.5 \cdot 0.4 + 0.5 \cdot 0.4}{1 - 0.12} = 0.68 \tag{2.9}$$

$$m(H_2) = \frac{0.5 \cdot 0.2}{1 - 0.12} = 0.11 \tag{2.10}$$

$$m(\Theta) = \frac{0.5 \cdot 0.4}{1 - 0.12} = 0.23 \tag{2.11}$$

2.3.2.3 Application de la théorie de l'évidence au groupement perceptuel

Cadre de discernement

Dans notre cas, on dispose grâce au graphe d'adjacence des régions, d'un certain nombre d'hypothèses de groupement. Pour chaque arête du graphe, on considère alors un cadre de discernement composé de deux hypothèses : $\Theta = \{G_{ij}, \overline{G_{ij}}\}$ avec :

- G_{ij} qui désigne l'hypothèse *grouper la région i et la region j*
- $\overline{G_{ij}}$ représente l'hypothèse *ne pas grouper la région i et la region j.*

Jeux de masses

On dispose alors pour chaque cadre de discernement Θ de plusieurs jeux de masses, représentant chacun une propriété Gestalt. On considère alors que chaque jeu de masses engage une croyance sur deux évènements seulement :

- G_{ij} d'une part (hypothèse simple)
- l'incertitude résiduelle d'autre part : Θ (hypothèse composée : $\Theta = G_{ij} \cup \overline{G_{ij}}$)

La figure 2.15(a) montre un exemple d'un tel jeu de masses.

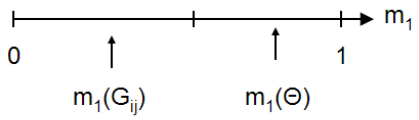


FIG. 2.15 – Exemple de jeu de masses utilisé pour le groupement perceptuel : la croyance se répartit sur G_{ij} et Θ.

Renforcement de la croyance par combinaison des jeux de masses

Il faut bien noter que, du fait de nos éléments focaux, il n'y a pas de conflit entre nos hypothèses durant la combinaison ($k = 0$, voir la figure 2.16). Ainsi, les résultats de l'équation (2.6) seront toujours représentatifs.

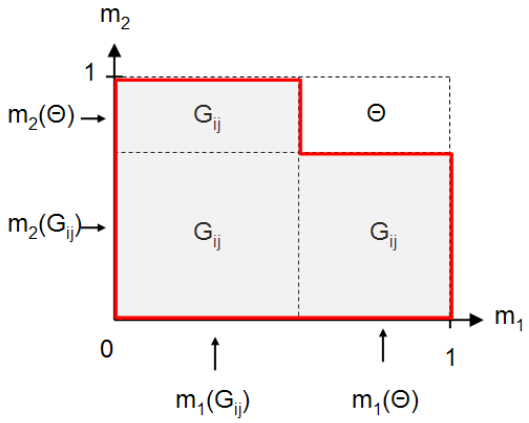


FIG. 2.16 – Combinaison de deux jeux de masses issus de la figure 2.15.

Cette modélisation est très différente de celle qu'en aurait fait des probabilités classiques. En effet, dans ce cas, la croyance aurait été répartie entre les deux hypothèses G_{ij} et $\overline{G_{ij}}$. Ainsi, un conflit entre les propriétés aurait pu apparaître et entraîner une impossibilité de conclure quant à la pertinence d'un groupement.

En application de la règle de Dempster (2.6), on a alors, pour une combinaison de deux jeux de masses m_1 et m_2 :

$$m(G_{ij}) = m_1(G_{ij}) + m_2(G_{ij})(1 - m_1(G_{ij})) \tag{2.12}$$

La figure 2.17 représente $m(G_{ij})$ comme une fonction des deux variables $m_1(G_{ij})$ et $m_2(G_{ij})$.

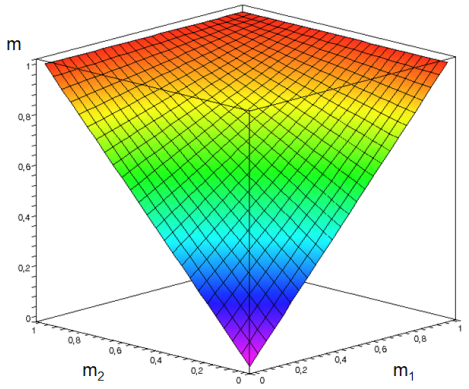


FIG. 2.17 – Représentation graphique de $m(G_{ij})$ comme une fonction des deux variables $m_1(G_{ij})$ and $m_2(G_{ij})$.

On peut remarquer que :

$$m(G_{ij}) > m_1(G_{ij}) \text{ et } m(G_{ij}) > m_2(G_{ij}) \tag{2.13}$$

Ceci signifie que les croyances en une hypothèse de groupement G_{ij} ont tendance à se renforcer sous l'influence conjointe des masses $m_k(G_{ij})$. Ainsi, lorsque plusieurs propriétés Gestalt sont activées sur une hypothèse de groupement, elles agissent de manière coopérative en faveur de cette dernière. Des exemples numériques de différentes combinaisons sont présentés plus loin.

Généralisation à n jeux de masses

Notre but est de combiner plus de deux jeux de masses sur un même cadre de discernement, puisque, rappelons-le, un jeu de masses modélise l'influence d'une propriété Gestalt sur une hypothèse de groupement. Les résultats énoncés plus haut peuvent être généralisés pour un ensemble de n jeux de masses, en appliquant itérativement l'équation (2.6). En effet, puisque nous combinons des jeux de masses sur des cadres de discernement identiques, la théorie de l'évidence nous indique que la règle de Demspter s'applique et que le jeu de masses résultant existe toujours.

2.3.3 Description des propriétés Gestalt

Cette section détaille les descripteurs utilisés pour quantifier les propriétés Gestalt sur chacune des hypothèses. Il n'y a pas de consensus, même parmi les psychologues fondateurs de l'Ecole Gestalt (KOFFKA, 1935; KANIZSA, 1979), sur le nombre

de propriétés à l'oeuvre durant la perception, ni sur leur aspect computationnel. Des exemples de ces propriétés ont été présentés dans la figure 2.2.

Nous avons proposé un certain nombre de descripteurs pour rendre compte de quelques-unes de ces propriétés, qui nous paraissaient particulièrement importantes : la proximité, la similarité, la fermeture et une propriété mixte de continuité/parallélisme. Notons que la propriété de proximité est directement prise en compte par l'utilisation d'un graphe d'adjacence pour générer les hypothèses de groupement.

2.3.3.1 Notion de descripteurs

Nous cherchons ici à obtenir, pour chaque propriété Gestalt, et pour chaque cadre de discernement (arête du graphe d'adjacence), une mesure brute M_k qui rende compte de l'influence de la propriété sur le groupement potentiel. Ces mesures brutes seront ensuite normalisées par la notion de saillance afin de fournir des jeux de masses m_k.

Dans un souci d'harmonisation, les mesures brutes M_k ont été formalisées de manière à les faire tendre vers 0 lorsque la propriété Gestalt correspondante est réalisée. Elles représentent donc des distances perceptuelles, entre les régions R_i et R_j engagées.

Enfin, précisons tout de suite que les implémentations proposées sont guidées par des considérations de performance en temps de calcul, puisque nous cherchons à les utiliser dans une approche d'indexation.

2.3.3.2 Propriété de similarité

La notion de similarité est pensée du point de vue des descripteurs utilisés lors de la phase de sur-segmentation, préalable à la génération du graphe d'adjacence des régions. Ainsi, l'image à traiter se compose de régions R_i, homogènes en terme d'un certain descripteur bas-niveau (couleur, texture) $d_{i,k}$. Nous proposons donc de définir une mesure de similarité M_1 comme une distance euclidienne entre ces descripteurs :

$$M_1(G_{ij}) = \left(\sum_k (d_{i,k} - d_{j,k})^2 \right)^{1/2} \tag{2.14}$$

En particulier, nous avons utilisé par la suite une sur-segmentation couleur par l'algorithme mean-shift de Comaniciu et Meer (1997). Nous utilisons donc pour $d_{i,k}$ un descripteur couleur dans l'espace L,a,b, car une distance euclidienne dans cet espace rend compte explicitement de la différence perceptuelle pour l'humain entre les couleurs engagées.

La mesure $M_1(G_{ij})$ tend donc vers 0 lorsque les jeux de descripteurs $d_{i,k}$ des régions considérées sont identiques, perceptuellement. Elle tend donc à favoriser le groupement de deux régions ayant des descripteurs bas-niveaux proches.

2.3.3.3 Propriété de fermeture

D'après Koffka (1935), de l'Ecole Gestalt, la fermeture tend à privilégier la perception d'objets simples, aux contours fermés, d'apparence compacte. Pour mesurer une telle propriété, hautement subjective, nous modélisons chaque région R_i par son ellipse englobante E_{R_i}. Cette dernière est définie commme l'ellipse qui possède les même moments d'ordre 2 que la région. Ainsi :

$$M_2(G_{ij}) = \left| 1 - \frac{\text{aire}(R_i + R_j)}{\text{aire}(E_{R_i+R_j})} \right| \tag{2.15}$$

où $R_i + R_j$ représente la région issue de la fusion des régions R_i et R_j. Nous utilisons une ellipse englobante afin d'approximer rapidement en temps de calcul une enveloppe convexe.

Ainsi, $M_2(G_{ij})$ tend vers 0 lorsque la région $R_i + R_j$ prend la forme d'une ellipse. Ce descripteur favorise donc la création d'objets arrondis ou la fusion d'une région fortement imbriquée dans une autre.

2.3.3.4 Propriété de continuité / parallélisme

Au contraire des deux propriétés précédentes, la continuité et le parallélisme reposent bien plus sur les contours des régions considérées que sur les régions en elle-mêmes. Nous avons choisi de les unifier dans une propriété unique, car elles reposent toutes deux sur l'orientation relative des segments composant les contours. La différence principale entre les deux provient du fait que la continuité se joue pour deux segments proches alors que le parallélisme impose à ces derniers d'être distants.

Nous utilisons donc une approximation polygonale des contours des régions, basée sur l'approximation récursive de Pavlidis et Horowitz (1974). Ensuite, l'orientation θ_{s_m} de chaque segment s_m est calculée d'après un axe de référence (voir figure 2.18).

La mesure brute de la propriété de continuité/parallélisme se base alors sur la différence d'orientation des segments considérés. On pose donc :

$$M_3^*(G_{ij}) = \min_{(s_m,s_n)\in(\mathcal{S}_i\times\mathcal{S}_j)} |\theta_{s_m} - \theta_{s_n}| \tag{2.16}$$

avec $\mathcal{S}_i$ qui représente l'ensemble des segments issus de l'approximation polygonale de la région R_i. Des exemples de structures détectées pour cette propriété sont

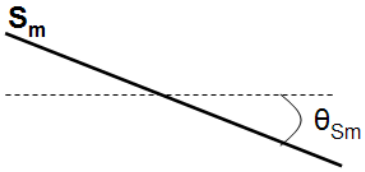


FIG. 2.18 – Orientation θ_{s_m} d'un segment s_m.

présentés dans la figure 2.19.

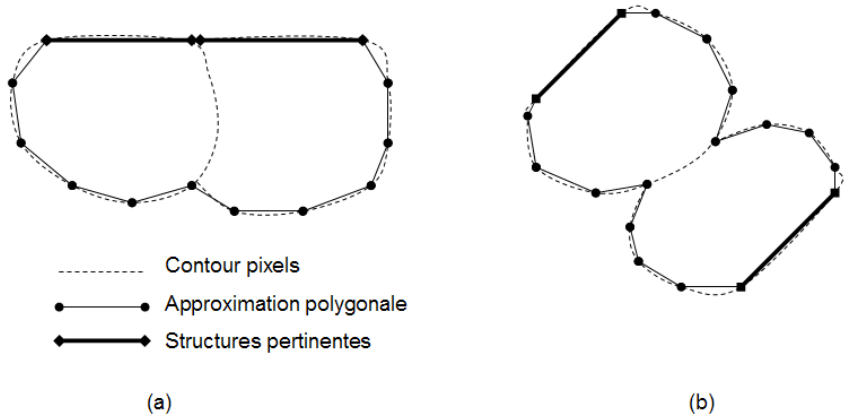


FIG. 2.19 – Exemple de structures détectées pour la propriété de continuité / parallélisme. En (a) continuité des frontières. En (b) parallélisme des contours.

Deux paramètres correctifs sont appliqués à la mesure brute $M_3^*(G_{ij})$: α_{s_m} et $\beta_{s_m s_n}$, qui empêchent la détection de structures comme celles de la figure 2.20. En (a) les segments considérés ont la même orientation mais ne sont pas représentatifs de leur région. Ils ne peuvent donc pas activer la propriété. En (b) bien que les segments engagés soient chacun représentatif de leur région, ils n'ont pas une taille comparable.

Ces paramètres correctifs sont calculés de la manière suivante :

$$\alpha_{s_m} = \frac{\max\limits_{s_k \in \mathcal{S}_i}(l_{s_k})}{l_{s_m}} \qquad \beta_{s_m s_n} = \frac{\max(l_{s_m}, l_{s_n})}{\min(l_{s_m}, l_{s_n})} \tag{2.17}$$

avec l_{s_m} la longueur du segment s_m. Noter que α_{s_n} est le terme analogue de α_{s_m} pour la région R_j. Noter également que l'on a $\alpha_{s_m} > 1$ et $\beta_{s_m s_n} > 1$.

Finalement, la mesure de la propriété mixte de continuité parallélisme est obtenue par :

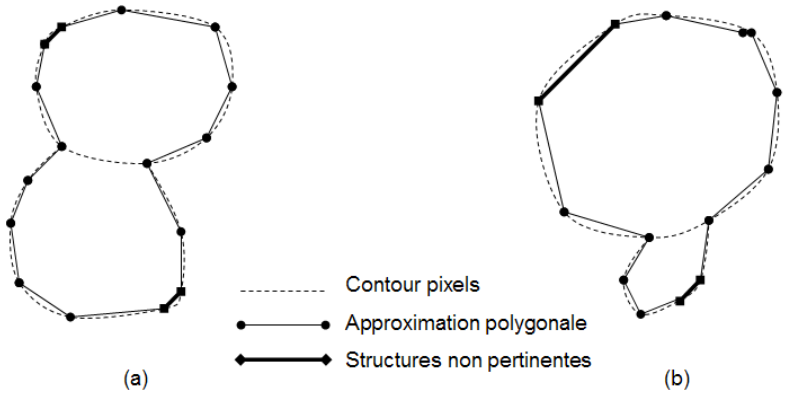


FIG. 2.20 – Exemple de structures non pertinentes pour la propriété de continuité / parallélisme.

$$M_3(G_{ij}) = \min_{(s_m,s_n)\in(\mathcal{S}_i\times\mathcal{S}_j)} (|\theta_{s_m} - \theta_{s_n}|\, \alpha_{s_m}\alpha_{s_n}\beta_{s_m s_n}) \tag{2.18}$$

2.3.4 Evaluation de la saillance des propriétés

A cette étape, nous avons besoin d'un processus de normalisation des mesures brutes $M_k(G_{ij})$ afin d'en déduire les jeux de masses m_k : à partir d'une mesure de propriété, comment en déduire la confiance que cette propriété va porter sur le groupement correspondant ?

Nous avons choisi de dériver cette confiance en fonction de la saillance de la mesure brute. En effet, nous avons déjà vu que la perception favorise les stimuli en fort contraste avec leur environnement. Nous cherchons donc à rendre la confiance portée sur un groupement d'autant plus forte que le descripteur correspondant est saillant. Il faut bien noter que nous sommes ici dans un contexte différent de celui évoqué par OLIVA ET AL. (2003), qui évaluait la saillance d'un élément pixel. Ici, nous évaluons la saillance d'une mesure de propriété sur un groupement.

Nous avons donc modélisé chaque mesure brute M_k comme une variable aléatoire qui suit une loi gaussienne, de densité de probabilité f_k :

$$f_k = \frac{1}{\sqrt{2\pi}} e^{\left(-\frac{M_k-\mu_k}{2\sigma_k}\right)^2} \tag{2.19}$$

En notant $p(M_k)$, la probabilité de la mesure brute M_k, sur toute l'image, la saillance S du descripteur k correspondant vaut alors :

$$S(M_k) = \frac{1}{p(M_k)} \tag{2.20}$$

D'où on en déduit la confiance m_k accordée à l'hypothèse de groupement G_{ij} :

$$m_k(G_{ij}) = \frac{2}{N}\left(1 - \frac{S(\mu_k)}{S(M_k)}\right) = \frac{2}{N}\left(1 - \frac{p(M_k)}{p(\mu_k)}\right) \qquad \text{si } M_k \leq \mu \tag{2.21}$$

$$m_k(G_{ij}) = 0 \qquad \text{si } M_k > \mu \tag{2.22}$$

Ainsi, si pour une hypothèse donnée, un descripteur M_k prend une valeur moyenne(μ_k), le rapport $\frac{p(M_k)}{p(\mu_k)}$ vaudra 1 d'où une croyance associée nulle : une valeur "moyenne" de descripteur n'est pas saillante et ne conduit à aucune croyance pour le groupement.
Au contraire, la croyance est maximale lorsque $p(M_k)$ diminue, c'est-à-dire lorsque M_k s'éloigne de la moyenne.

N est le nombre de propriétés Gestalt utilisées. Le facteur de normalisation $\frac{2}{N}$ assure ainsi qu'aucune propriété ne peut à elle seule assurer une confiance totale (1) en une hypothèse de groupement. Au contraire, elle a besoin de la coopération d'un certain nombre d'autres propriétés pour ceci.

Enfin, par souci de normalisation, au sens de Dempster-Shafer, on pose également :

$$m_k(\Theta) = 1 - m_k(G_{ij}) \tag{2.23}$$

Ceci signifie qu'une propriété Gestalt ne peut allouer de croyance que sur deux évènements (voir à ce sujet la partie modélisation 2.3.2) :

- l'hypothèse simple G_{ij} qui stipule que les régions R_i et R_j seront regroupées
- l'hypothèse composée Θ qui rend compte de l'incertitude résiduelle

2.4 Resultats

2.4.1 Resultats sur des images artificielles

Tout d'abord, un certain nombre de tests ont été menés sur des images artificielles et ce, afin de vérifier l'exactitude des implémentations proposées des différentes propriétés Gestalt. Ceci permet également de bien illustrer le mécanisme de coopération des propriétés. Nous présentons maintenant une série de résultats, liés à l'image de la

figure 2.21. Plusieurs hypothèses de groupement vont être générées à partir de cette image, en particulier, quatre hypothèses relatives au regroupement des différentes structures principales.

FIG. 2.21 – Image artificielle de test.

2.4.1.1 Evaluation des descripteurs liés aux propriétés Gestalt

Nous allons maintenant détailler les résultats sur chacune de ces hypothèses de groupement. Pour plus de clarté, chaque hypothèse sera entourée en rouge sur la figure présentée. Pour chacune, trois croyances partielles sont affichées. Elles correspondent aux croyances engagées par les trois propriétés Gestalt de similarité, fermeture et continuité/parallélisme. Rappelons qu'une croyance est, de manière générale, un nombre réel compris entre 0 (pas de croyance) et 1 (certitude absolue). Ici, du fait de la normalisation des descripteurs, chaque croyance est bornée entre 0 (pas de croyance) et $0,66$ (croyance partielle totale).

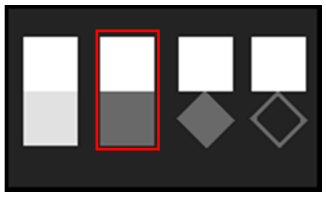
Le premier test est présenté en figure 2.22. Dans ce cas, on observe que les trois propriétés Gestalt sont activées, puisqu'elles ont toutes les trois une croyance partielle associée très forte (rappelons que ces dernières sont bornées à $0,66$). En effet, la propriété de similarité a été activée par la faible distance entre niveaux de gris des deux carrés. La fermeture a également bien répondu du fait de la création d'un rectangle, forme compacte, lors de la fusion éventuelle des deux carrés. La règle de continuité / parallélisme a, elle, été déclenchée pour deux raisons : le parallélisme entre côtés opposés du carré d'une part, et la continuité entre les côtés adjacents des carrés d'autre part.

La figure 2.23 présente un cas de figure identique au précédent, à la différence des niveaux de gris des deux carrés concernés. Cette fois, cet écart de niveau de gris est, visuellement, plus important. On constate alors que la propriété de similarité est beaucoup moins activée que précédement (croyance partielle de $0,13$). Les autres croyances partielles sont, elles, inchangées, ce qui est conforme à l'intuition.

Dans l'exemple de la figure 2.24, on a considéré la fusion éventuelle d'un losange et d'un carré. Cette fois, aucune continuité ni parallélisme n'apparaît entre les deux

$m_1(G_{ij}) = 0.58$ (croyance de similarité)
$m_2(G_{ij}) = 0.58$ (croyance de fermeture)
$m_3(G_{ij}) = 0.66$ (croyance de continuité/parallélisme)

FIG. 2.22 – Evaluation des descripteurs des propriétés Gestalt (test 1).



$m_1(G_{ij}) = 0.13$ (croyance de similarité)
$m_2(G_{ij}) = 0.58$ (croyance de fermeture)
$m_3(G_{ij}) = 0.66$ (croyance de continuité/parallélisme)

FIG. 2.23 – Evaluation des descripteurs des propriétés Gestalt (test 2).

figures géométriques, et il en résulte une croyance partielle très faible par rapport à la propriété de continuité/parallélisme.
Notons également que le regroupement éventuel de ces deux figures géométriques (carré et losange) créerait un objet moins compact que les objets considérés précédement (rectangles). On constate ainsi une croyance partielle relative à la propriété de fermeture significativement inférieure aux exemples 2.22 et 2.23 $(0, 36)$.

$m_1(G_{ij}) = 0.13$ (croyance de similarité)
$m_2(G_{ij}) = 0.36$ (croyance de fermeture)
$m_3(G_{ij}) = 0.20$ (croyance de continuité/parallélisme)

FIG. 2.24 – Evaluation des descripteurs des propriétés Gestalt (test 3).

La figure 2.25 illustre le cas limite où aucune propriété ne favorise le groupement. En effet, à la différence de l'exemple précédent (figure 2.24), le losange considéré est cette fois creux. Il en résulte que le regroupement éventuel de cet objet avec le carré au-dessus créerait un objet très peu compact, puisque creux. On observe alors une croyance partielle relative à la propriété de fermeture très faible $(0, 06)$.

Ainsi, ces exemples nous permettent de constater que notre implémentation des descripteurs des différentes propriétés Gestalt semble correcte.

$m_1(G_{ij}) = 0.13$ (croyance de similarité)
$m_2(G_{ij}) = 0.06$ (croyance de fermeture)
$m_3(G_{ij}) = 0.20$ (croyance de continuité/parallélisme)

FIG. 2.25 – Evaluation des descripteurs des propriétés Gestalt (test 4).

2.4.1.2 Evaluation de la combinaison des descriteurs liés aux propriétés Gestalt

Nous présentons maintenant, pour chaque hypothèse, en plus des croyances partielles relatives à chaque propriété Gestalt, une croyance totale, qui correspond à la combinaison des trois croyances partielles. Rappelons qu'à la différence des croyances partielles, à cause du processus de normalisation, la croyance totale est comprise entre 0 (pas de croyance) et 1 (certitude).

Le premier test est présenté en figure 2.26.

$m_1(G_{ij}) = 0.58$ (croyance de similarité)
$m_2(G_{ij}) = 0.58$ (croyance de fermeture)
$m_3(G_{ij}) = 0.66$ (croyance de continuité/parallélisme)

$m(G_{ij}) = 0.94$ (croyance totale)

FIG. 2.26 – Evaluation de la combinaison des descripteurs des propriétés Gestalt (test 1).

Comme chacune des trois propriétés (similarité de niveau de gris, fermeture, continuité/parallélisme) a répondu favorablement au groupement, on constate que la croyance total au groupement est très élevée. Ceci est dû au mécanisme de renforcement mutuel de croyance entre les propriétés, dont nous avons déjà parlé. A titre de comparaison, la figure 2.27 présente un autre résultat dans lequel, cette fois, la propriété de similarité ne favorise pas le groupement (croyance partielle de 0.13). Bien que les autres croyances partielles restent élevées, la croyance totale est significativement inférieure au test précédent.

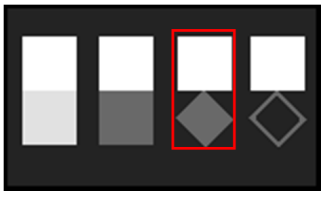
Toujours dans le même ordre d'idée, le test de la figure 2.28 illustre le cas où une seule propriété active le groupement : la propriété de fermeture, tendant à la création d'objets fermés et compacts étant assez importante, la croyance totale demeure significative.

Il faut remarquer à ce stade que même lorsque certaines propriétés ne soutiennent pas explicitement le groupement, le système reste capable de conclure, sans s'arrêter à un simple conflit des propriétés entre elles.

$m_1(G_{ij}) = 0.13$ (croyance de similarité)
$m_2(G_{ij}) = 0.58$ (croyance de fermeture)
$m_3(G_{ij}) = 0.66$ (croyance de continuité/parallélisme)

$m(G_{ij}) = 0.87$ (croyance totale)

FIG. 2.27 – Evaluation de la combinaison des descripteurs des propriétés Gestalt (test 2).



$m_1(G_{ij}) = 0.13$ (croyance de similarité)
$m_2(G_{ij}) = 0.36$ (croyance de fermeture)
$m_3(G_{ij}) = 0.20$ (croyance de continuité/parallélisme)

$m(G_{ij}) = 0.55$ (croyance totale)

FIG. 2.28 – Evaluation de la combinaison des descripteurs des propriétés Gestalt (test 3).

La figure 2.29 illustre le cas limite où aucune propriété ne favorise le groupement. Dans ce cas, la croyance totale demeure bien inférieure à tous les autres cas précédents.

$m_1(G_{ij}) = 0.13$ (croyance de similarité)
$m_2(G_{ij}) = 0.06$ (croyance de fermeture)
$m_3(G_{ij}) = 0.20$ (croyance de continuité/parallélisme)

$m(G_{ij}) = 0.34$ (croyance totale)

FIG. 2.29 – Evaluation de la combinaison des descripteurs des propriétés Gestalt (test 4).

2.4.2 Réduction du graphe

2.4.2.1 Principe de la réduction

Etant donnée, pour chaque arête du graphe d'adjacence, une valeur de croyance totale au groupement, il est possible de réduire ce graphe itérativement. Pour l'instant, nous utilisons un algorithme glouton pour cette réduction, qui fusionne à chaque étape les deux noeuds séparés par l'arête qui possède la plus forte valeur de confiance au groupement. Cette méthode nous assure que le processus de réduction converge. Néanmoins, il peut éventuellement se trouver piégé dans un minimum local. Nous verrons néanmoins qu'un tel algorithme de réduction donne déjà des

résultats très satisfaisants.

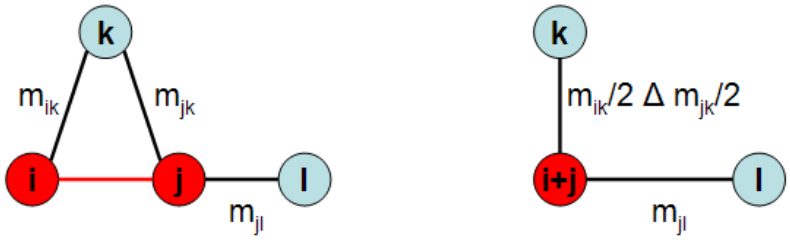
Ainsi, le graphe est réduit jusqu'à ce qu'il ne comporte plus aucune arête dont la confiance au groupement ne soit supérieur à une valeur donnée (notée minBelief). Cette dernière est directement reliée à la granularité du groupement. Plus elle est élevée, et moins les régions de l'image seront regroupées. L'initialisation de cette valeur par un utilisateur est donc assez intuitive. Nous verrons toutefois dans le chapitre suivant qu'il est possible, et même souhaitable, de s'affranchir d'une telle valeur.

2.4.2.2 Mise en oeuvre

En pratique, il est important de noter que les mesures de propriétés Gestalt ne sont calculées qu'une seule fois, avant le processus de réduction. Deux raisons majeures président à ce choix :

- La rapidité de traitement ainsi obtenue, pour chaque image. A titre indicatif, le temps de traitement d'une image naturelle de 256*384 pixels (sur-segmentation préalable, extraction des mesures de propriétés, combinaison et réduction du graphe) est ainsi de l'ordre de la seconde sur un Pentium IV à 3 GHz. Le temps de traitement varie en pratique en fonction du nombre de régions issu de la sur-segmentation préalable et donc de la complexité intrinsèque de l'image.
- Plus important, l'idée qui préside au groupement perceptuel est de faire émerger d'une image une série de zones saillantes. Nous considérons donc que la quantification des propriétés Gestalt et leur normalisation par saillance doit se faire par rapport à l'image originale et non pas itérativement. Si les propriétés sont requantifiées à chaque itération de réduction, on prend le risque de faire émerger de l'image des structures visuelles qui n'étaient pas saillantes au départ, mais qui le deviennent a posteriori.

En pratique, deux cas se produisent lors de la réduction du graphe, illustrés dans la figure 2.30(a). Dans cette dernière, les noeuds i et j sont en cours de fusion. Les deux cas possibles correspondent aux deux autres noeuds k et l.



(a) (b)

FIG. 2.30 – Exemple de graphe d'adjacence (a) et réduction correspondante (b).

Les noeuds i, j et k forment un sous-graphe complet

Dans ce cas, lors de la fusion des noeuds i et j, la question se pose de savoir quelle valeur de confiance transmettre entre le noeud k et le noeud résultant $i+j$. En effet, encore une fois se pose la question de la combinaison de deux mesures de confiance m_{ik} et m_{jk} en une seule valeur résultante. Nous utilisons à nouveau la théorie de l'évidence à cet effet. L'opérateur Δ de la figure 2.30 désigne donc la règle de combinaison de Dempster entre deux jeux de masses, allouant une croyance au groupement respectivement de $m_{ik}/2$ et $m_{jk}/2$.

La division par un facteur 2 est une normalisation préalable qui évite que la croyance d'un seul des arcs n'active à lui seul le groupement. Au contraire, la coopération des deux arcs est nécessaire au renforcement de la confiance combinée.

Les noeuds i, j et l ne forment pas un sous-graphe complet

Dans ce cas, la valeur de confiance m_{jl} entre les noeuds j et l est simplement transmise à l'arc entre le noeud l et le noeud résultant $i+j$.

2.4.3 Résultats sur des images naturelles

2.4.3.1 Stratégie de validation du groupement perceptuel : positionnement

Il est très difficile d'évaluer les résultats d'une segmentation ou d'un groupement perceptuel. En effet, ces traitements ne constituent pas une fin en soi, mais une étape indispensable à des traitements postérieurs. Martin et al. (2001) proposent un cadre d'évaluation, qui repose sur la comparaison des résultats avec des segmentations effectuées par des opérateurs humains.

Toutefois, ces segmentations manuelles demeurent assez subjectives et en tout cas totalement déconnectées de traitements postérieurs. Ainsi, une image donnée peut être vue comme sous-segmentée, ou, au contraire, comme sur-segmentée, suivant le but recherché. Par exemple, si nous sommes intéressés par l'extraction des objets par rapport au fond des images, les résultats sur les poissons de la figure 2.31 sont excellents. Si maintenant nous cherchons à comparer chaque poisson extrait avec des modèles stockés dans une base, en prenant en compte une dimension structurelle (position des sous-parties), ces résultats deviennent sous-segmentés. Encore une fois, nous ne pourrons évaluer quantitativement nos résultats qu'une fois revenus à des cas d'utilisation d'indexation. Cette évaluation *globale* du système sera proposée dans la section 3.5 du chapitre suivant.
Il faut rappeler que nous ne cherchons pas une nouvelle méthode de segmentation ou de groupement perceptuel robuste. Au contraire, nous pensons qu'il n'en n'existe pas dans un univers non contraint. Notre objectif est de déployer une chaine complète de

traitements pour l'indexation et la requête aux bases d'images. Nous verrons par la suite que certains traitement permettent de corriger certaines erreurs du groupement perceptuel, lors de la comparaison entre une requête et une image de la base. Ce n'est donc qu'une fois toute la chaine considérée, que nous pourrons pointer les éventuels problèmes liés au groupement perceptuel, dans *notre* perspective d'indexation.

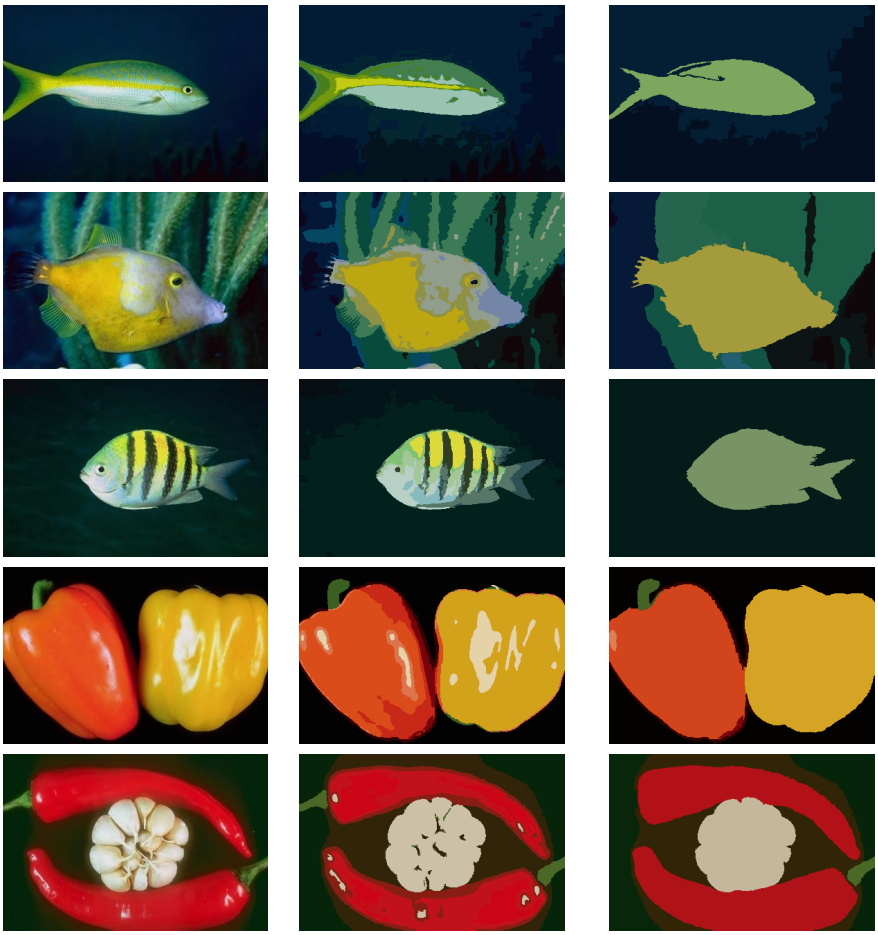
Nous nous contenterons donc pour l'instant d'une évaluation qualitative du groupement perceptuel. Nous avons déjà montré dans la partie précédente que l'implémentation proposée des propriétés Gestalt semblait conforme à l'intuition et que le mécanisme de coopération des différentes lois était pertinent. Nous allons maintenant comparer qualtitativement notre système à d'autres systèmes existants, sur une base d'images de référence (Corel©).

2.4.3.2 Exemple de résultats sur des images naturelles

Nous avons testé notre processus de groupement perceptuel sur un échantillon d'images issues de la base bien connue Corel©. Des exemples de résultats sont montrés à la figure 2.31.

On peut constater que le groupement perceptuel est capable de réduire de manière significative le bruit induit par l'étape de segmentation : les nombreuses régions de très petite taille sont fusionnées de manière à créer des régions de taille significative. En outre, il a tendance à faire émerger de l'image des zones qui correspondent aux différents objets sémantiques (par exemple, les poissons dans les trois premières images et les poivrons, piments des deux autres images).

On peut également remarquer que les artefacts d'éclairage induits par la segmentation sont corrigés de manière assez efficace. Ainsi, les reflets sur les poivrons et piments des deux dernières images sont éliminés lors du groupement perceptuel, en regroupant les régions correspondantes dans l'objet proprement dit. La propriété de fermeture joue ici un rôle particulièrement important.



(a) image originale (b) pré-segmentation (c) groupement perceptuel

FIG. 2.31 – Exemple de résultats de groupement perceptuel (minBelief respectivement à 50%, 40%, 35%, 62%, 55%).

D'autres résultats sont présentés dans les figures 2.32 et 2.33. Comme les images sont cette fois plus complexes, la phase de groupement perceptuel ne parvient pas à extraire totalement les différents objets sémantiques. Toutefois, des structures pertinentes, correspondant aux différentes parties des objets en question, émergent : les roues et la carrosserie des voitures (figure 2.32) ainsi que la peau et les vêtements des personnages (figure 2.33). En outre, les différentes régions qui composent les fonds d'image à l'issue de la segmentation préalable sont bien regroupées entre elles, ce qui permet d'isoler correctement le sujet de son fond.

On peut noter par exemple l'utilisation conjointe des propriétés de similarité et de parallélisme pour extraire le sol de la deuxième image de la figure 2.32, ainsi que la carrosserie de la voiture dans la première image.

(a) image originale (b) pré-segmentation (c) groupement perceptuel

FIG. 2.32 – Exemple de résultats de groupement perceptuel (minBelief respectivement à 43%, 68%, 70%, 62%).

Concernant la figure 2.33, les cheveux du personnage de la 3ème image sont déjà regroupés avec une partie du fond de l'image dès l'étape de segmentation préalable. Comme le groupement perceptuel s'appuie sur cette description, il ne peut isoler les cheveux du fond. C'est pourquoi ces derniers sont entièrement fusionnés avec le fond de l'image.

(a) image originale (b) pré-segmentation (c) groupement perceptuel

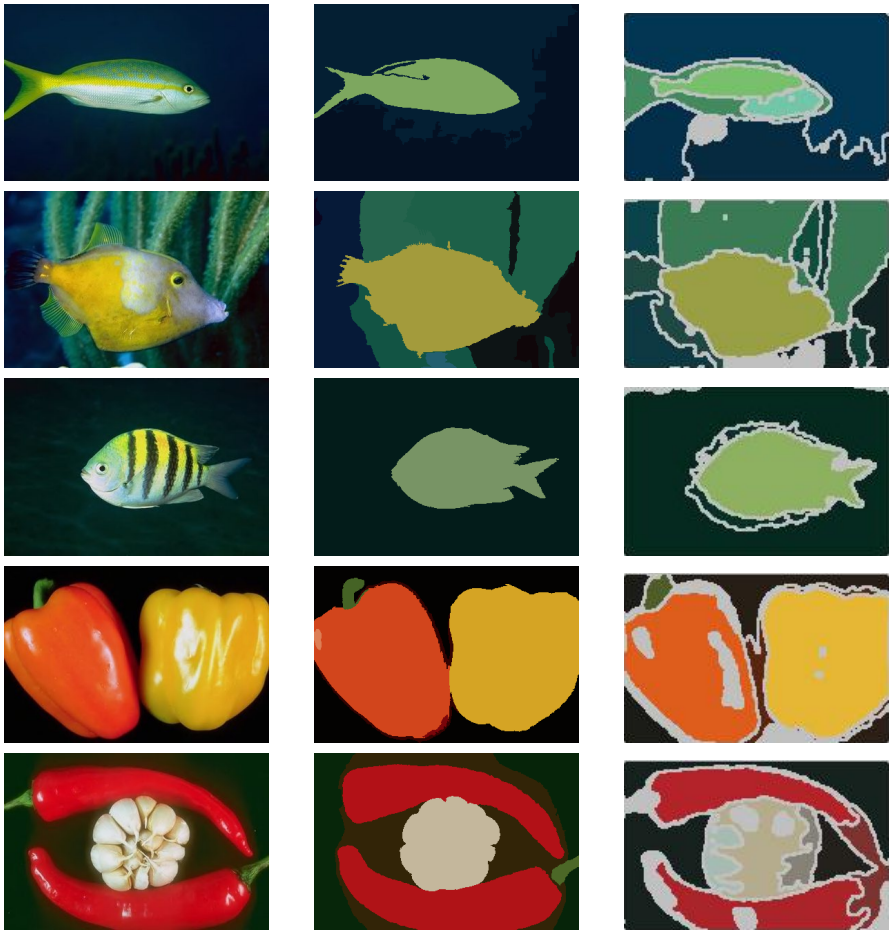
FIG. 2.33 – Exemple de résultats de groupement perceptuel (minBelief respectivement à 55%, 50%, 71%).

2.4.3.3 Comparaison à d'autres systèmes

Comparaison à une segmentation

Rappelons que la segmentation peut être vue comme la première étape d'un groupement perceptuel. Il est donc naturel, dans un premier temps, de se comparer à des systèmes de segmentation. La figure 2.34 présente ainsi les résultats obtenus par le système Blobworld, dont nous avons déjà parlé. Rappelons que ce dernier opère un regroupement de pixels sur la base de trois critères : proximité spatiale,

similarité de couleur et de texture. Les pixels affichés en gris correspondent à des zones de l'image qui ont été éliminées des traitements.



(a) image originale (b) groupement perceptuel (c) Blobworld

FIG. 2.34 – Comparaison du groupement perceptuel au système Blobworld.

On constate que le groupement perceptuel permet de fusionner des régions qui ont des descripteurs de couleur ou de texture différents, mais qui présentent tout de même une certaine unité. Ainsi, les poissons sont correctement extraits du fond par le groupement perceptuel, alors que la segmentation par Blobworld les laissent partitionnés en plusieurs régions.

On remarque aussi que, du fait des limitations inhérentes à la segmentation, les résultats de Blobworld présentent des artefacts dûs aux conditions d'éclairage (sur les poivrons et les piments). Au contraire, le groupement perceptuel permet de supprimer ces derniers, grâce aux propriétés de continuité/parallélisme et surtout de fermeture.

Comparaison à d'autres systèmes de groupement

D'autres comparaisons ont été effectuées, avec des systèmes explicitement de groupement perceptuel. Les résultats sont présentés à la figure 2.35. Deux systèmes sont considérés à titre comparatif. Le premier (colonne (c)) a été proposé par Idrissi et al. (2004) et consiste à réduire itérativement un graphe d'adjacence en utilisant pour chaque arête un score de groupement. Ce dernier est basé sur une similarité couleur, pondérée par un paramètre rendant compte de la propriété de fermeture. Bien que des résultats de qualité aient été obtenus avec ce système, la surimportance donnée à la similarité couleur conduit à des erreurs de groupement, comme dans la troisième image ou le corps du personnage est fusionné avec le fond. Un autre exemple de telles erreurs est visible sur la deuxième image, dans laquelle le bras du personnage est fusionné avec son genou.
En outre, les résultats obtenus avec ce système sont souvent sur-segmentés par rapport aux notres. Ceci vient du fait que le processus de groupement d'Idrissi et al. (2004) nécessite souvent d'être stoppé rapidement, en vue d'éviter l'apparition d'erreurs. Au contraire, comme notre système impose à plusieurs propriétés Gestalt d'interagir pour activer un groupement, il est capable d'effectuer plus de groupements avant l'apparition d'erreurs. Ceci permet un gain certain en robustesse.

(a) image originale | (b) groupement perceptuel | (c) Idrissi et al. (2004) | (d) Luo et Guo (2003)

Fig. 2.35 – Comparaison du groupement perceptuel à d'autres systèmes.

Le second système considéré est celui de Luo et Guo (2003) (colonne (d)). Il est basé sur une modélisation par champs de Markov (MRF). On constate que les résultats sont d'une qualité comparable aux notres. Par exemple, si notre système ne parvient pas à extraire les cheveux du personnage dans la troisième image, tout le visage est regroupé en une seule région, à la différence du résultat de Luo et Guo (2003).

Notons toutefois que les auteurs utilisent, pour caractériser chaque hypothèse de groupement, une probabilité d'apparition obtenue avec une somme pondérée d'éléments qui correspondent à chacune des propriétés Gestalt. Or, l'initialisation de ces paramètres se fait de manière empirique et peut donc vite devenir hasardeuse pour un utilisateur. Au contraire, notre système, qui repose sur une normalisation

de chaque propriété par saillance, ne propose qu'un seul paramètre ($minBelief$), relié à la granularité de la description souhaitée. Il est donc assez facile à initialiser de manière intuitive. En outre, on peut remarquer que sa plage de variation, pour obtenir les résultats présentés, est assez étroite : elle varie entre 40 et 70%.

2.5 Conclusion sur le groupement perceptuel

Nous avons présenté dans ce chapitre un nouveau système de groupement perceptuel dans un contexte pré-attentif. Il s'appuie sur une sur-segmentation préalable en régions, à l'issue de laquelle un graphe d'adjacence est extrait. Rappelons que cette sur-segmentation intègre une information sur les contours extraits de l'image, afin de limiter les artefacts inhérents aux méthodes régions.

Combiner des points de vue par la théorie de l'évidence

Le graphe d'adjacence obtenu sert de support pour générer différentes hypothèses de groupement entre régions. Chacune de ces hypothèses est caractérisée quantitativement selon trois propriétés inspirées de la théorie Gestalt. Nous utilisons alors la théorie de l'évidence de Dempster Shafer afin de pouvoir combiner l'influence de ces propriétés, de manière à déduire pour chaque hypothèse une valeur unique caractérisant sa probabilité d'apparition. Ce formalisme nous permet donc de modéliser finement l'interaction souhaitée entre les propriétés Gestalt. Dans un souci de robustesse, nous imposons à plusieurs (mais pas nécessairement toutes) propriétés de soutenir le groupement, afin d'activer ce dernier.

Aller au-delà de la segmentation

Les mécanismes de groupement perceptuel permettent d'améliorer significativement les résultats d'une segmentation couleur classique. Ainsi, il est possible de fortement réduire le nombre de régions décrivant une image et de faire émerger des structures saillantes, correspondant aux objets contenus dans les images. En outre, plusieurs artefacts induits par la segmentation peuvent être corrigés, comme ceux dûs aux conditions d'éclairage.

De la pertinence d'une description multiple

A ce stade, le groupement perceptuel nous a permis de donner une description unique d'une image donnée. Or, nous avons déjà fait remarquer que le résultat attendu de tels traitements varie en fonction du but recherché. Par exemple, le niveau de détail du résultat est variable selon les structures à extraire de l'image. La perspective d'utiliser le groupement perceptuel dans un système d'indexation pose

donc problème, puisqu'on ne peut pas savoir a priori ce qu'il est pertinent d'indexer. Il est donc difficile, voire impossible de déterminer correctement le paramètre d'arrêt du groupement perceptuel ($minBelief$).

C'est la raison pour laquelle nous ne chercherons pas à indexer une image par une description unique. Au contraire, nous l'indexerons par une série de descriptions, qui correspondent aux différentes étapes du groupement perceptuel. Ainsi, une image est décrite par différents niveaux de groupements perceptuels, à différents niveaux de détail. Lors de la requête, selon le type d'objet cherché ou le niveau de détail souhaité, on se placera à un niveau différent de la hiérarchie de groupement. La figure 2.36 présente, à titre d'illustrations, les principales étapes d'un groupement perceptuel. On y voit l'émergence progressive des sous-parties de l'objet principal (marteau), puis de l'objet en tant que tel.

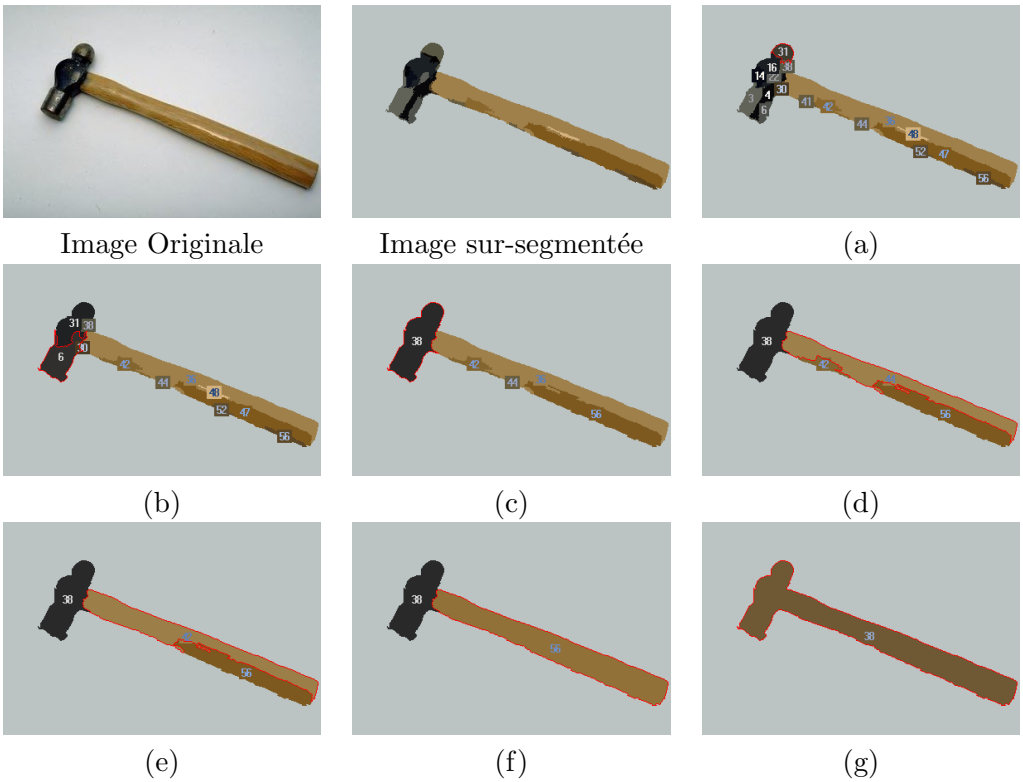


FIG. 2.36 – Principales étapes d'un groupement perceptuel (a-g) à partir d'une image originale.

Qu'en est-il des traitements attentifs ?

Il est évident que dans la hiérarchie de groupement vont se trouver un certain nombre de régions sans importance sémantique aucune. Néanmoins, et encore une

fois, il est difficile de faire un tri pertinent a priori, lors de l'indexation. En revanche, lors de la requête, selon le type d'objet souhaité, il sera possible de mettre en place une série de traitements (dits attentifs cette fois, puisque pilotés par des connaissances externes) qui auront pour but d'interpréter les structures extraites par le groupement perceptuels, et éventuellement de corriger celle-ci, ou d'en invalider certaines. C'est ce que nous nous proposons de décrire dans le prochain chapitre.

3

Du groupement perceptuel à l'indexation structurelle

Sommaire

3.1 Introduction

3.1.1 Retour sur la problématique de l'indexation

Nous avons vu dans le chapitre 1 que les systèmes d'indexation existants permettent généralement de rechercher des images d'après leurs caractéristiques bas-niveaux comme la couleur ou la texture. Forsyth et al. (1997) remarquent que ceci autorise l'utilisateur à formuler des requêtes orientées *matière*, *matériau* (le mot anglais utilisé est *stuff*). Or, si cet aspect est sans nul doute pertinent, l'utilisateur cherche néanmoins le plus souvent une image d'après ce qu'elle représente, en particulier les *objets* qu'elle représente. Ainsi, un utilisateur aimerait formuler une requête du type : *je recherche les images qui contiennent des voitures.*

Ceci conduit à deux remarques. D'une part, les outils d'indexation se doivent d'intégrer des caractéristiques image qui permettront de se rapprocher du niveau sémantique des images. En effet, nous l'avons déjà vu, le fossé sématique empêche un tel lien d'être direct et non ambigu. Toutefois, l'utilisation de descriptions structurelles permettent par exemple une première extension du pouvoir descriptif des données image.

D'autre part, un système d'indexation doit offrir aux utilisateurs une interface capable de les aider à formuler au mieux ce qu'ils recherchent. Le paradigme désormais classique de la requête par l'exemple, largement diffusé et reconnu, permet de retrouver des images jugées similaires à une image requête. Néanmoins, ce paradigme reste limité puisqu'il est incapable de généraliser, à partir de l'exemple fourni, ce que cherche véritablement l'utilisateur.

Durant ces travaux de thèse, nous avons proposé un nouveau paradigme d'indexation, qui prend en compte ces deux constatations. Un système a été developpé, en application : il s'agit du moteur THIS, acronyme anglais de *THings-oriented Image retrieval System*. Ce chapitre est dédié à sa présentation.

3.1.2 Positionnement

3.1.2.1 Vers une hiérarchie de segmentations

Nous l'avons vu, la question de la segmentation d'image est une question centrale en indexation. Elle est presque impossible à réaliser dans des domaines non contraints, sans assistance de l'utilisateur. Ainsi, elle conduit souvent à des descripteurs extraits sur des régions non pertinentes. Par exemple, lorsqu'un objet comporte une ombre, la segmentation va conduire à séparer la zone d'ombre de l'objet et fournira alors un jeu de descripteurs pour chacune d'entre elles. La caractérisation de l'objet en tant que tel est alors manquante.

Néanmoins, la segmentation permet de manipuler un certain nombres de des-

cripteurs, en particulier ceux reliés à la notion de forme, qui se révèle être particulièrement expressifs pour caractériser des objets du monde réel. De nombreuses expériences ont montré la pertinence de tels descripteurs sur des bases d'objets présegmentés. C'est en partant de cette constatation que nous avons choisi d'utiliser une segmentation pour extraire nos descripteurs. Toutefois, avertis des limites inhérentes à un tel procédé, nous proposons deux extensions :

- Nous utilisons d'une part une notion plus générale que la segmentation : le groupement perceptuel, présenté dans le chapitre 2.
- D'autre part, conscients que même cette notion généralisée sera entachée d'erreurs, nous ne nous appuyons pas sur *un* résultat de segmentation, mais plutôt sur *des* segmentations, à travers une *hiérarchie*. Celle-ci décrit une image à différents niveaux de détails : depuis une vision fine à une description très grossière. Comme nous le verrons par la suite, le fait d'utiliser cette hiérarchie de segmentations permet de corriger certaines des erreurs liées à celle-ci. En outre, cette structure de données se prête très naturellement à l'utilisation de descripteurs structurels, augmentant encore le pouvoir expressif de la signature.

3.1.2.2 Vers une requête orientée *modèle*

Concernant maintenant le second point, à savoir le type d'interface de requête, nous proposons une interrogation par modèle. Plus précisément, dans le cas où l'utilisateur recherche une image d'après un objet qu'elle représente, il nous semble pertinent de recourir à un prototype. Nous entendons par prototype une entité abstraite, formalisée par l'utilisateur, qui "résume" synthétiquement ce qu'il cherche.

L'importance de cette notion de prototype a déjà été mise en évidence par les expériences de Tversky (1977). Son utilisation procure deux avantages notoires. Tout d'abord, il donne la possibilité à l'utilisateur, lors de son élaboration, de réfléchir plus précisément à ce qu'il cherche : quelles caractéristiques sont indispensables, lesquelles ne sont qu'optionnelles, par exemple. Ainsi, à partir d'une idée souvent vague, l'élaboration du prototype permet d'affiner la requête souhaitée.
Ensuite, le prototype permet de synthétiser les caractéristiques essentielles de l'objet cherché. Ainsi, dans le cas d'une requête par l'exemple, il peut y avoir un décalage entre ce que cherche réellement l'utilisateur et l'image requête fournie : peut-être l'utilisateur souhaitait-il une voiture de profil, alors que l'image requête est légèrement décalée. La conception d'un prototype permet de clarifier de genre d'ambiguités, en allant à l'essentiel.

Nous présentons dans la section suivante l'architecture générale du système, avant de détailler les caractéristiques utilisées dans la partie 3.3. La section 3.4 présente ensuite comment nous utilisons la théorie de l'évidence à deux niveaux différents pour combiner différents descripteurs lors de la requête. Enfin, différents résultats sont proposés dans la section 3.5.

3.2 Architecture générale

3.2.1 Définitions et notations

3.2.1.1 Notion de modèle

On appelle *modèle* la requête formulée par l'utilisateur. Un modèle M est constitué d'un objet, lui-même composé de différentes parties (sous-parties du modèle), notées $M_1, \ldots, M_n$. Chaque partie M_i est caractérisée par différents descripteurs, comme la forme ou ses relations spatiales par rapport à l'objet entier. Ces descripteurs seront présentés dans la section 3.3.

On suppose qu'un objet peut être reconnu par sa liste non ordonnée de sous-parties. La figure 3.1 présente un exemple de modèle.

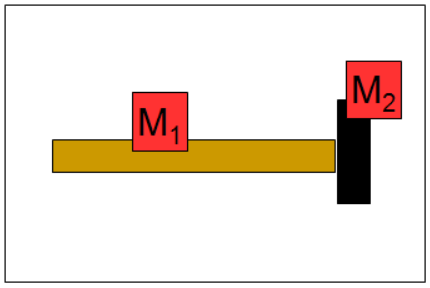


FIG. 3.1 – Exemple de requête (modèle) composée de deux parties M_1 et M_2.

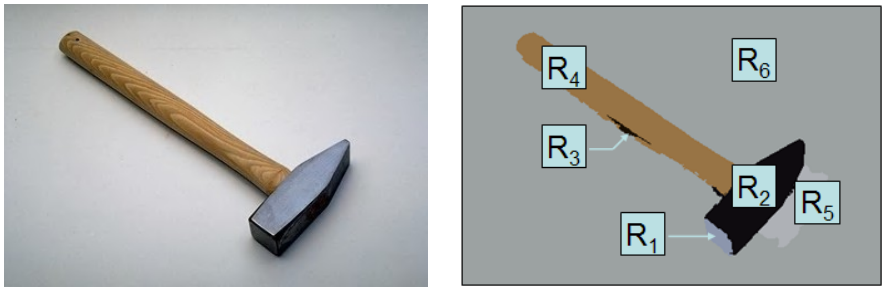
3.2.1.2 Notion d'arbre de régions

On appelle *arbre de régions* $R_1, \ldots, R_m$, la hiérarchie de régions issue du groupement perceptuel présenté dans le chapitre précédent : lorsque deux régions R_i et R_j sont groupées dans une troisième R_k, les noeuds correspondant R_i et R_j de l'arbre sont placés comme fils du noeud R_k. Un arbre de régions décrit ainsi une image à différents niveaux de détail. La figure 3.3 présente un exemple d'arbre de régions, obtenu à partir de la sur-segmentation de la figure 3.2.

L'arbre de régions possède ainsi six feuilles, correspondant aux six régions de la sur-segmentation : $R_1, \ldots, R_6$. On peut remarquer sur cet exemple, que le groupement perceptuel commet une erreur en groupant la tête du marteau (R_7) avec son ombre portée sur le fond (R_5). Ainsi, le noeud résultant R_8 comporte une erreur.

3.2.1.3 Notion de noeuds sémantiques

On appelle *noeud sémantique* un noeud de l'arbre de régions qui correspond à un objet du monde réel. L'exemple d'arbre de régions de la figure 3.3 possède ainsi



(a) image (b) sur-segmentation couleur

FIG. 3.2 – Exemple d'image et de sur-segmentation couleur associée.

deux noeuds sémantiques : R_7 et R_9, correspondant respectivement à la tête et au manche du marteau.

L'objet du système d'indexation sera de reconnaître un noeud sémantique, quelque soit sa position a priori dans l'arbre. Ceci procure une souplesse d'utilisation bien plus grande que la segmentation, dans laquelle chaque noeud sémantique aurait dû être extrait comme *résultat* du traitement. Ainsi, bien que la tête du marteau ait été groupée avec son ombre portée dans la figure 3.3 (noeud R_8), elle peut toujours être reconnue car elle apparaît plus bas dans l'arbre, au noeud R_7.

3.2.2 Une requête : apparier un modèle et un arbre de région

Dans ce contexte, traiter une requête à une base d'images consiste à rechercher un modèle M dans un arbre de régions R. Pour ceci, nous cherchons à calculer une distance[1] globale $|M - R|$. Nous proposons pour ceci un processus à trois étapes, illustré dans la figure 3.4.

3.2.2.1 Recherche d'un sous-arbre optimal

Tout d'abord, nous recherchons un *sous-arbre optimal* ST_k, qui contient tous les noeuds sémantiques et le moins d'autres noeuds possible. Ce sous-arbre est en fait la meilleure approximation de l'objet cherché, extrait de son fond. Comme nous le verrons par la suite, la connaissance de cette approximation est nécessaire pour calculer les descripteurs structurels. Dans la figure 3.4, ST_{10} représente le sous-arbre optimal.

Puisqu'il n'existe pas de méthode robuste pour extraire un tel sous-arbre optimal, nous considérons tous les sous-arbres possibles ST_k pour calculer la distance globale

[1]Nous utilisons le mot *distance* abusivement. Comme nous le verrons par la suite, cette fonction de similarité ne vérifie pas les axiomes d'une distance

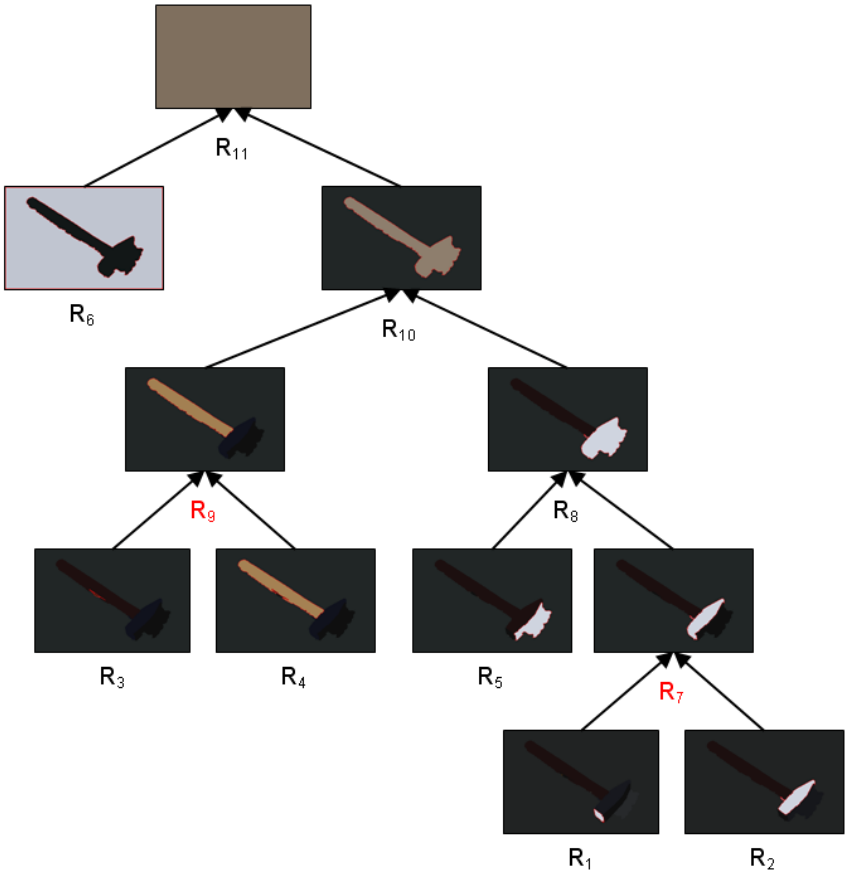


FIG. 3.3 – Exemple d'arbre de régions, obtenus à partir de la sur-segmentation de la figure 3.2

.

$|M - R|$. A la fin des traitements, nous conservons le meilleur, c'est-à-dire celui qui minimise la distance $|M - R|$.

3.2.2.2 Appariement régions / partie de modèle

Pour chaque sous-arbre ST_k, nous cherchons à apparier chaque partie M_i du modèle avec la région R_j de ST_k qui correspond le mieux. Pour évaluer cette correspondance, nous combinons plusieurs descripteurs (forme, structure) afin de déduire une distance unique $|M_i - R_j|_k$ pour chaque appariement potentiel entre M_i et R_j.

Notons que cette distance dépend du sous-arbre ST_k considéré.

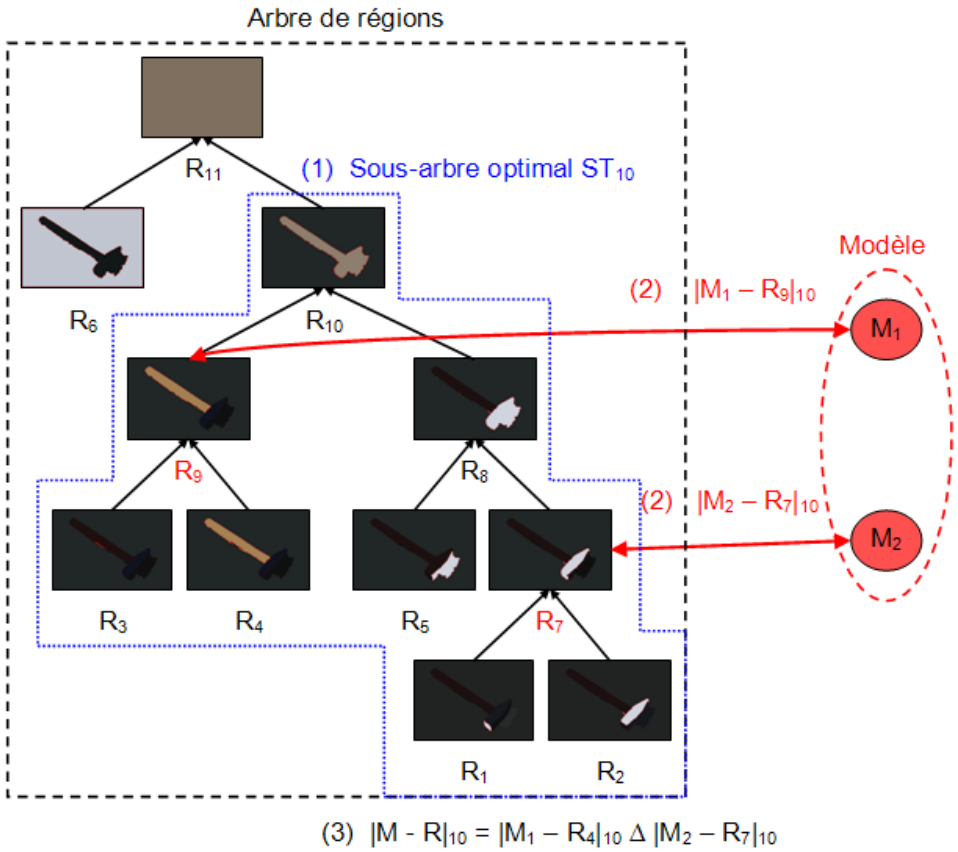


FIG. 3.4 – Recherche de modèle M dans un arbre de régions.

3.2.2.3 Appariement global sous-abre optimal / modèle

Chacune des distances $|M_i - R_j|_k$ sert à l'évaluation d'une distance globale $|M - R|_k$ entre le modèle entier et le sous-arbre considéré ST_k (3). Cette étape de combinaison de distances est notée Δ dans la figure 3.4. Elle met en jeu la théorie de l'évidence et sera décrite dans la section 3.4.

3.2.3 Apparier une région et une partie de modèle

Nous l'avons vu, l'étape (2) du processus consiste à rechercher autant d'appariements possibles entre les parties M_i du modèle et les régions R_j du sous-arbre ST_k, tout en minimisant les distances entre elles : $|M_i - R_j|_k$. Il faut bien voir que cette question ne peut se résoudre avec les algorithmes classiques du couplage maximum entre graphes, car dans notre cas, tous les ancêtres et fils d'un noeud déjà apparié ne peuvent plus être appariées à leur tour. La figure 3.5 présente un exemple d'illustration, dans lequel la partie M_1 du modèle (manche du marteau) vient d'être

appariée avec la région R_9. Dans ce cas, la partie restante du modèle (M_2) ne peut plus être appariée ni avec les régions filles de R_9 (R_3 et R_4) ni avec les régions mères du sous-arbre considéré (R_{10}).

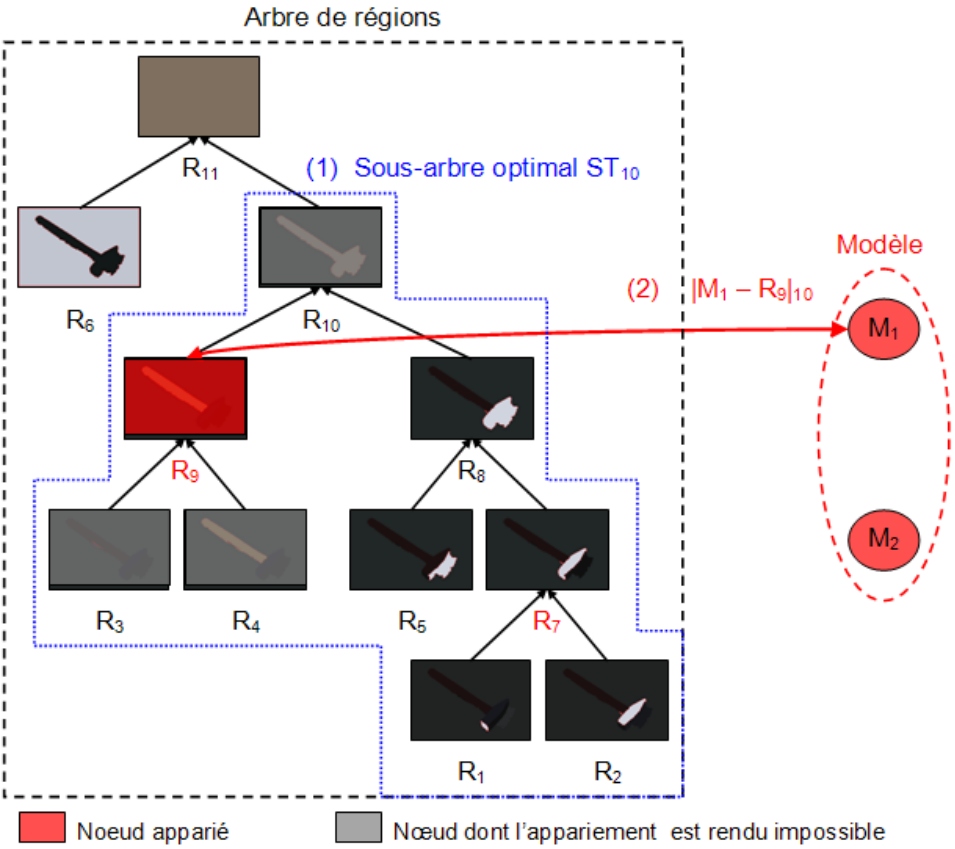


FIG. 3.5 – Exemple d'appariement région-parties rendus impossibles, en raison d'appariement préalables.

Ainsi, au lieu de tester tous les cas d'appariement possibles, nous utilisons un algorithme glouton, qui se révèle être suffisant pour nos besoins. Il consiste à calculer toutes les distances $|M_i - R_j|$ possibles, et à apparier alors le couple M_i, R_j qui minimise la distance $|M_i - R_j|$. Une fois éliminées les régions mères et filles de la région appariée, le procédé est ré-itéré jusqu'à l'appariement de toutes les parties M_i du modèle, ou des régions R_j du sous-arbre considéré.

3.3 Descripteurs utilisés

Nous allons maintenant présenter les descripteurs utilisés afin de comparer une région R_j de l'arbre de région aux parties M_i du modèle. On distingue deux types

de descripteurs :

- ceux qui dépendent uniquement de la région traitée, que nous appellerons descipteurs région ;
- ceux qui dépendent de la région traitée ainsi que de l'objet référence auquel elles appartiennent, c'est-à-dire du sous-arbre optimal considéré. Nous qualifierons ces descripteurs de structurels.

3.3.1 Descripteurs régions

Nous avons considéré que lors de la comparaison de régions significatives, issues de l'arbre de régions, avec une partie de modèle, le descripteur le plus pertinent a trait à la forme. Bien que des descripteurs plus bas-niveau, comme la couleur ou la texture soient utiles dans un premier temps, lors d'une sur-segmentation par exemple, ils ne sont pas robustes en terme de reconnaissance d'objets. En effet, ils comportent trop d'artefacts du fait de leur trop grande sensibilité aux conditions d'éclairement.

Comme nous l'avons vu dans la section 1.3.2.2, il existe un très grand nombre de descripteurs de forme. De nombreux travaux (Zhang et Lu, 2004), parmi lesquels ceux du comité MPEG-7 (Zaharia et Prêteux, 2004) ont constaté que l'utilisation conjointe de descripteurs basés régions et contours se révèle être particulièrement efficace. Rappelons qu'un descripteur basé région décrit une forme par sa distribution spatiale de pixels, alors qu'une méthode basée contour ne considère que le contour de l'objet pour caractériser celui-ci.

Nous avons donc utilisé les descripteurs ART (*Angular Radial Transform*) et CSS (*Curvature Scale Space*) comme descripteurs de formes basés région et contour respectivement.
ART consiste en une transformation complexe définie sur un disque unité en coordonnées polaire (Kim et Kim, 1999). C'est un descripteur robuste à l'échelle, à la rotation et à la translation.
CSS (Mokhtarian et Mackworth, 1992) est un descripteur de forme basé contour. Comme ART, il est robuste au changement d'échelle, à la rotation et à la translation. Il consiste à suivre les positions des points d'inflexions d'un contour, lorsque celui-ci est soumis à une série de filtres gaussiens de largeur variable. Lorsque la largeur augmente, les inflexions peu marquées sont éliminées et le contours devient plus lisse. Les points d'inflexions qui subsistent sont considérés comme étant caractéristiques du contour.

3.3.2 Descripteurs structurels

La plupart des systèmes d'indexation considère les objets comme un tout, sans prendre en compte d'information structurelle. Pourtant, celle-ci peut se révéler forte-

ment utile. Par exemple, si un utilisateur recherche des drapeaux d'un certain type, la répartition spatiale de ses composants est beaucoup plus importante que l'objet dans sa globalité, dont la forme rectangulaire est très peu discriminante. Dans cette optique, nous avons choisi de modéliser trois descripteurs de structures, pour chaque région :

- la position relative de la région par rapport à l'objet entier.
- la taille relative de la région par rapport à l'objet entier.
- l'orientation relative de la région par rapport à l'objet entier.

Il faut noter que toutes ces caractéristiques sont *relatives* à l'objet entier. Elles dépendent donc du sous-arbre ST_k choisi, qui représente l'objet dans sa globalité. Ainsi, nous avons besoin d'une estimation grossière de la taille et de l'orientation de cette zone de l'image. Pour ceci, nous utilisons l'ellipse englobante E_k du sous-arbre ST_k en cours de traitement.
Lorsque nous souhaitons apparier la région R_j d'un sous-arbre ST_k à une partie M_i de modèle, nous commençons par recaler le sous-arbre ST_k sur le modèle M, grâce à leurs ellipses englobantes. Le processus est illustré dans la figure 3.6. On calcule une transformation affine T qui recale l'ellipse englobante E_k de ST_k sur l'ellipse englobante E du modèle M. Ensuite, T est appliquée sur l'ellipse englobante E_j de la région R_j. On compare alors $T(E_j)$ et l'ellipse englobante E_i de M_i.

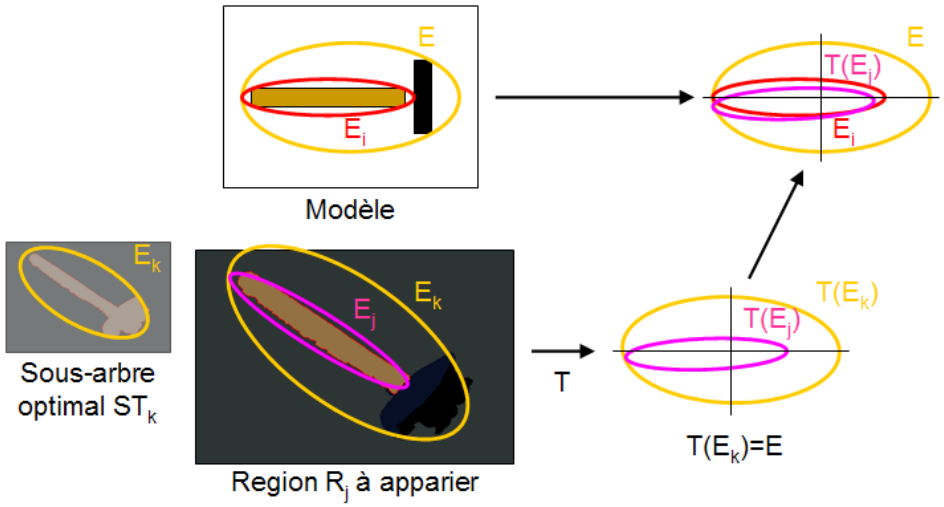


FIG. 3.6 – Recalage d'une région sur un modèle, pour l'extraction des descripteurs structurels.

Plus précisément, la comparaison structurelle comporte trois descripteurs :

- la distance euclidienne entre les centroïdes des ellipses $T(E_j)$ et E_i ;
- le rapport d'aires entre $T(E_j)$ et E_i ;
- la différence d'orientation entre $T(E_j)$ et E_i ;

3.4 Combinaison des descripteurs par la méthode de l'évidence

Lorsque l'on compare une région R_j avec une partie M_i de modèle, nous prenons en compte plusieurs descripteurs : les formes basées contours et régions, ainsi que les positions, tailles et orientations relatives à l'objet entier. Se pose alors, comme durant le groupement perceptuel, la question de la combinaison de ces descripeurs et ce, afin d'en déduire une valeur unique qui quantifiera la qualité de l'appariement. Nous appliquons donc encore une fois le formalisme de l'évidence de Shafer (1976), qui permet de combiner des points de vue différents sur une hypothèse. Ici, le formalisme est appliqué à deux niveaux :

- lors de l'appariement d'une région R_j avec une partie M_i du modèle d'une part
- ensuite, lors de l'appariement entre un sous-arbre ST_k et le modèle dans sa globalité M.

Dans les deux cas, chaque descripteur peut être vu comme un point de vue différent sur l'hypothèse d'appariement. La théorie de l'évidence nous permet alors de dériver un point de vue combiné sur l'appariement, qui synthétise tous les autres.

Nous ne reviendrons pas sur les bases de la théorie de l'évidence qui ont déjà été présentées dans la section 2.3.2.1. Il convient désormais de présenter l'utilisation que nous faisons de cette théorie lors de la requête. Notons que la notion de croyance de la théorie de l'évidence est strictement équivalente à celle de distance utilisée jusqu'à maintenant. En effet, une forte croyance correspond à une faible distance et réciproquement.

3.4.1 Combinaison des descripteurs lors de l'appariement région / partie

Pour chaque appariement potentiel entre une région R_j et une partie M_i de modèle, on considère le cadre de discernement suivant : $\Theta = \left\{O_{ij}, \overline{O_{ij}}\right\}$, avec :

- O_{ij} qui représente l'évènement : *la région R_j est appariée à la partie M_i du modèle*. R_j est donc reconnue comme une partie d'un *objet*, dont le modèle est M.
- $\overline{O_{ij}}$ qui représente l'évènement : *la région R_j n'est pas appariée à la partie M_i du modèle* ;

3.4.1.1 Jeux de masses utilisés

Cinq jeux de masses sont utilisés, représentant cinq points de vue sur ce cadre de discernement. Ils correspondent respectivement aux descripteurs ART, CSS, position relative, taille relative et orientation relative. Chacun de ces jeux de masses ne peut

engager une croyance que sur deux évènements : O_{ij} et Θ. Cela signifie qu'aucun descripteur ne peut être interprété directement de manière à nier l'appariement. Au contraire, un appariement doit, afin d'être déclenché, obtenir une croyance positive forte de plusieurs (mais par forcément tous) descripteurs. La figure 3.7 illustre un tel jeu de masses.

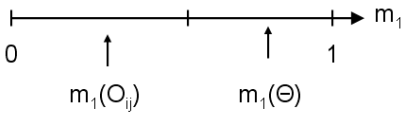


FIG. 3.7 – Jeu de masses utilisé pour l'appariement région/partie.

3.4.1.2 Des distances aux jeux de masses

Avant de rentrer dans le détail de la combinaison des jeux de masses, nous allons succintement présenter le mécanisme qui nous permet de transformer les distances entre descripteurs en croyance. Compte-tenu du rapport distance élevée - croyance faible (et réciproquement), nous proposons la forme :

$$m_p(O_{ij}) = \frac{1}{N} \exp^{\frac{-D_p}{\alpha_p}} \tag{3.1}$$

avec :

- O_{ij} désigne l'hypothèse d'appariement concernée, entre une région R_j du sous-arbre et une partie M_i du modèle.
- D_p désigne une distance entre une région R_j du sous-arbre et une partie M_i du modèle, du point de vue d'un seul descripteurs p (ART, CSS, ou structure).
- α_p est un paramètre, relative au descripteur p, qui permet de modéliser la dynamique du descripteur, lors de sa conversion en croyance. Il influe sur l'écrasement de la courbe exponentielle.
- le paramètre de normalisation $\frac{1}{N}$ (N étant le nombre de descripteurs utilisés) assure qu'aucun descripteur ne peut, à lui seul, emporter une croyance totale sur l'appariement.

Une représentation graphique de cette équation est donnée en figure 3.8.

3.4.1.3 Combinaison des jeux de masses

La combinaison de deux de ces jeux de masses est illustrée sur la figure 3.9. Puisque chaque jeu de masses engage une croyance sur deux évènements, il en résulte

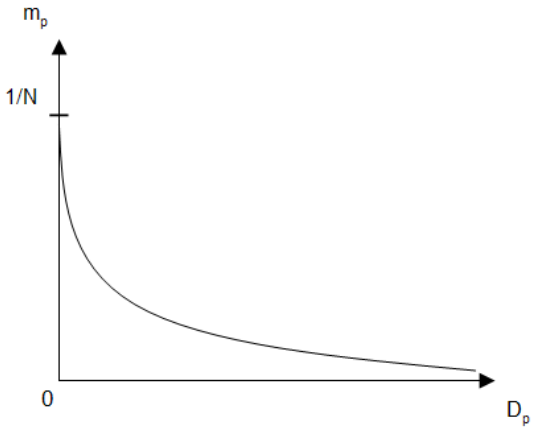


FIG. 3.8 – Equation de conversion distance à croyance.

quatre cas possibles. Parmi ceux-ci, trois soutiennent l'hypothèse O_{ij}. La croyance résultante en O_{ij} est alors représentée graphiquement par l'aire des trois zones correspondantes de la figure, entourée en rouge.

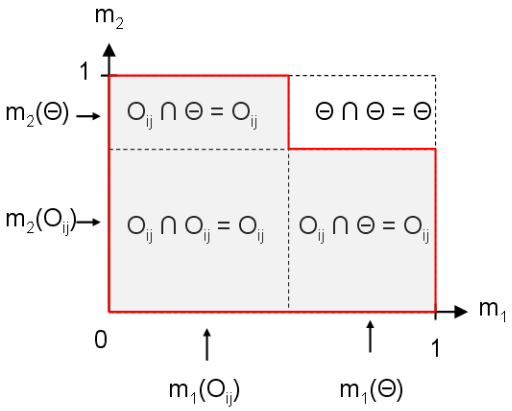


FIG. 3.9 – Combinaison de deux jeux de masses pour l'appariement région/partie (a).

On peut écrire alors, d'après la règle de combinaison de Dempster :

$$\begin{aligned} m(O_{ij}) &= m_1(O_{ij})m_2(O_{ij}) && (3.2)\\ &\quad + m_1(O_{ij})(1 - m_2(O_{ij})) && (3.3)\\ &\quad + m_2(O_{ij})(1 - m_1(O_{ij})) && (3.4)\\ &= m_1(O_{ij}) + m_2(O_{ij})(1 - m_1(O_{ij})) && (3.5) \end{aligned}$$

Du fait du cadre de discernement choisi, le conflit est nul entre les points de vue

($k = 0$) et les croyances en O_{ij} selon chaque descripteur ont tendance à se renforcer les unes avec les autres.

Un exemple de combinaison de deux jeux de masses, représentant les descripteurs ART et CSS, est présenté dans la figure 3.10. On constate que pour l'appariement du manche du marteau, les deux jeux de masses donnent une forte croyance en l'appariement et conduisent donc logiquement à une croyance combinée forte. Rappelons que les croyances allouées par les jeux de masses sont bornées (ici à 0.5), pour empêcher qu'un seul jeu (donc descripteur) n'emporte à lui seul une croyance totale. Pour l'appariement de la tête en revanche, le jeu de masses lié à ART alloue une forte croyance alors que celui de CSS répond un peu moins bien, du fait de la présence d'artefacts sur le contour de la région. Néanmoins, grâce à l'effet de renforcement, la croyance combinée est encore tout à fait acceptable, à la différence d'une valeur moyenne.

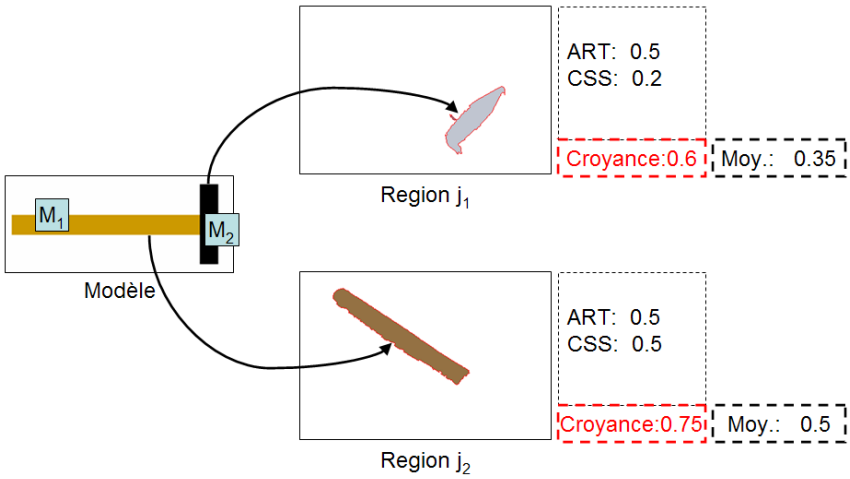


FIG. 3.10 – Exemple de combinaison de descripteur ART et CSS pour l'appariement région / partie.

A l'issue de ces traitements, on dispose donc d'un certain nombre d'appariements entre des régions R_j d'un sous-arbre ST_k avec des parties M_j de modèle. La pertinence de ces appariements est quantifiée par une croyance, notée $a_{ij,k}$. Cette dernière est un nombre réel, compris entre 0 (croyance nulle) et 1 (certitude). La situation est représentée à la figure 3.11.

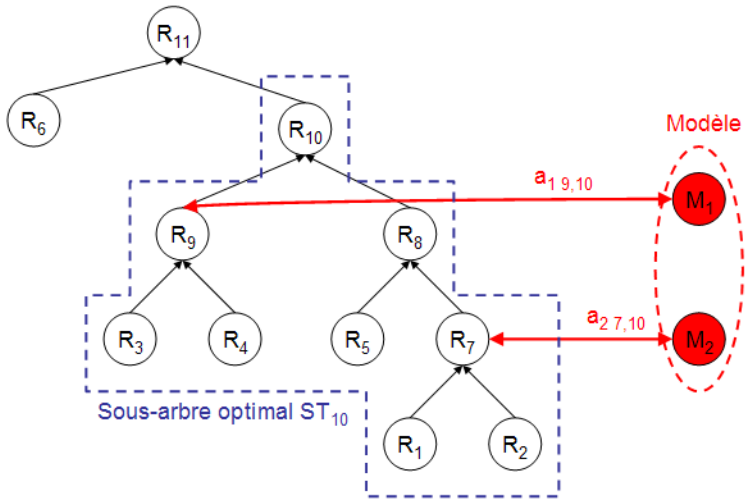


FIG. 3.11 – Etape de traitement après l'appariement région / partie.

3.4.2 Combinaison des descripteurs lors de l'appariement sous-arbre / modèle

Une fois que les régions R_j du sous-arbre ST_k ont été appariées aux parties M_i du modèle, nous calculons une distance globale entre le sous-arbre en question ST_k et le modèle entier M. Pour ceci, nous utilisons les scores d'appariement $a_{ij,k}$ de chacune des parties en considérant pour chaque partie M_i, le score $a_{ij,k}$ maximal.

Ainsi, pour chaque sous-arbre ST_k, nous considérons : $\Theta = \{O_k, \overline{O_k}\}$, avec :

- O_k qui représente l'évènement : *le sous-arbre ST_k est apparié au modèle M*. Ceci signifie que ST_k a été reconnu comme un *objet*, correspondant au modèle M.
- $\overline{O_k}$ qui représente l'évènement : *le sous-arbre ST_k n'est pas apparié au modèle M* ;

3.4.2.1 Jeux de masses utilisés

Contrairement au cas précédent, tous les jeux de masses utilisés ici n'ont pas la même structure.

Jeux de masses appuyant l'appariement

Intuitivement, chaque appariement région / partie entre une région R_j du sous-arbre ST_k et une partie M_i de modèle devrait contribuer à renforcement de l'appariement global entre le sous-arbre ST_k et le modèle M. C'est pourquoi chaque

appariement région / partie réalisé avec une croyance $a_{ij,k}$ conduit à la création d'un jeu de masses qui engage cette croyance $a_{ij,k}$ sur l'évènement O_k. Le reste de croyance $(1 - a_{ij,k})$ est alloué à l'incertitude Θ. Un tel jeu de masses est illustré dans la figure 3.12).

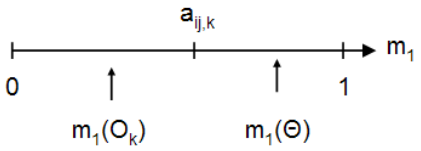


FIG. 3.12 – Jeu de masses utilisé pour appuyer l'appariement sous-arbre / modèle.

Jeux de masses pénalisant explicitement l'appariement

Nous introduisons également une pénalité à la croyance globale lorsqu'une partie M_i du modèle n'a pas été appariée dans le sous-arbre. En effet, cela signifie que le modèle n'a pas été entièrement reconnu. Cette pénalité prend la forme d'un jeu de masses qui engage une croyance $b_{ij,k}$ sur l'évènement $\overline{O_k}$. Ceci a pour effet de rejeter explicitement l'appariement, à la différence des autres cas précédents. Le reste de la croyance est toujours engagé sur Θ. Un tel jeu de masses est illustré dans la figure 3.13.

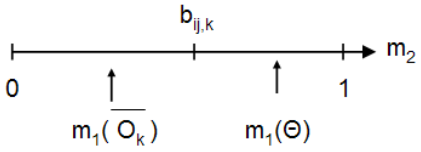


FIG. 3.13 – Jeu de masses utilisé pour pénaliser l'appariement sous-arbre / modèle.

En pratique, cette croyance $b_{ij,k}$ qui tend à rejeter l'appariement est fonction de l'aire de la partie M_i du modèle, non reconnue. Plus cette dernière est importante, et plus $b_{ij,k}$ aura tendance à rejeter l'appariement ($b_{ij,k} \rightarrow 1$).

3.4.2.2 Combinaison des jeux de masses

Cette fois, puisque tous les jeux de masses n'engagent pas une croyance sur des hypothèses similaires, le résultat de la combinaison est différent du cas précédent (évoqué au paragraphe 3.4.1). Trois situations peuvent apparaître, sans être mutuellement exclusives :

– la combinaison de deux jeux de masses soutenant l'hypothèse O_k (figure 3.12) conduit au renforcement de celle-ci et soutient donc la reconnaissance du modèle ;

- la combinaison de deux jeux de masses soutenant l'hypothèse $\overline{O_k}$ (figure 3.13) conduit au renforcement de celle-ci et contredit donc la reconnaissance du modèle ;
- la combinaison d'un jeu de masses soutenant O_k et d'un autre soutenant $\overline{O_k}$ conduit à quatre sous-cas, comme illustré dans la figure 3.14.

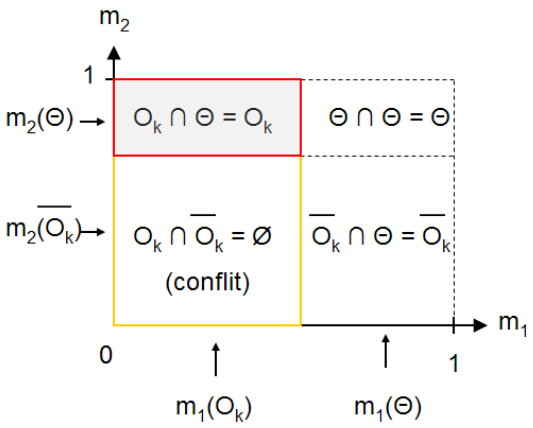


FIG. 3.14 – Combinaison de deux jeux de masses contradictoires pour l'appariement sous-arbre / modèle.

Dans cette dernière situation, une partie de la croyance se répartit sur les hypothèses O_k et $\overline{O_k}$. De plus, un conflit apparaît entre les points de vue. Ce dernier est quantifié par le nombre $k = m_1(O_k)m_2(\overline{O_k})$. Lorsque k tend vers 1, le conflit devient total, et la théorie de l'évidence nous recommande de ne pas conclure quant au résultat de la combinaison. Dans ce cas, nous posons une croyance globale en O_k nulle et le modèle n'est donc pas reconnu dans le sous-arbre ST_k.

Un exemple d'illustration est proposé dans la figure 3.15. Dans le cas (a), les deux parties du modèle ont bien été appariée à des régions de l'image. Il en résulte un renforcement de la croyance finale en la reconnaissance du modèle. En revanche, dans le deuxième cas (b), la tête du marteau n'a été appariée avec aucune région (croyance en appariement nulle). Il en résulte la création d'un jeu de masses qui soutient l'hypothèse de non reconnaissace $\overline{O_k}$ et conduit à une pénalité sur la croyance finale. La valeur de cette pénalité est dérivée de l'aire relative de la partie manquante.

A l'issue de ce processus, on dispose donc de deux croyances : une soutenant l'appariement, et l'autre pénalisant ce dernier. Tant que la pénalité reste inférieur à un seuil, nous ne considérons que la croyance soutenant l'appariement. Si cette pénalité dépasse le seuil autorisée (actuellement : 0.7), la croyance globale est remise à zéro. Il serait utile, dans des travaux futurs, de chercher des utilisations plus fine de ces deux types de croyance.

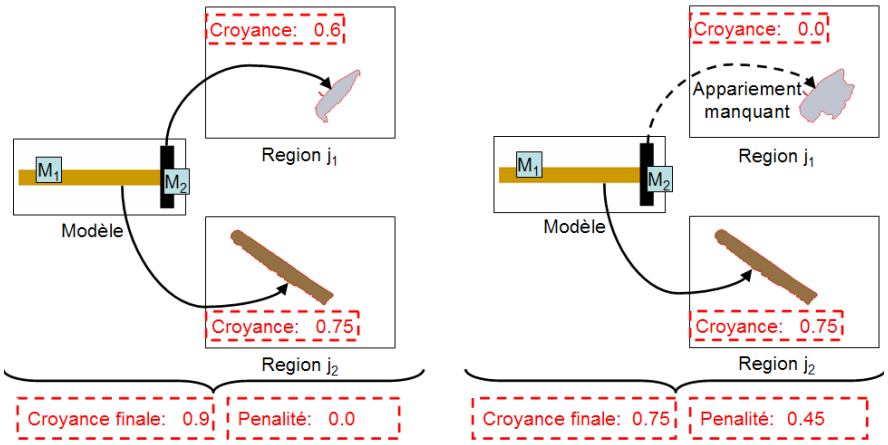


FIG. 3.15 – Exemple de combinaison de croyances pour l'appariement sous-arbre / modèle.

3.5 Résultats

Nous allons maintenant exposer différents résultats liés à cette partie d'indexation structurelle. Dans un premier temps (section 3.5.1), nous présenterons des résultats caractéristiques, sur un nombre réduit d'images, afin d'illustrer le comportement de notre système. Puis, nous présenterons les résultats de tests à grande échelle, sur une base de 600 images, dans la section 3.5.2.

Il est important de noter que notre système ne renvoie pas un résultat binaire tel que *cette image ne contient pas le modèle cherché*, mais au contraire quantifie la similarité mesurée entre le modèle et l'image. Cette similarité s'étend de 0 (le modèle n'a pas du tout été reconnu) à 1 (le modèle a parfaitement été reconnu).

3.5.1 Résultats caractéristiques

3.5.1.1 Modèle à deux parties

La figure 3.16 montre le résultat de la requête *marteau* sur un jeu réduit d'images. La modèle, présenté sur la première image à gauche, se compose de deux parties : la tête et le manche. Les valeurs numériques sous les images représentent la similarité mesurée entre l'image et le modèle. Rappelons que ces valeurs s'étendent de 0 (pas de similarité) à 1 (similarité totale). En outre, dans le cas d'une similarité non nulle, le système entoure les sous-parties reconnues par des ellipses rouges. L'ellipse jaune représente, elle, l'ellipse englobante de l'objet entier.
On constate ainsi que les images sans aucun marteau ont été jugées non similaires au modèle. Ceci montre que le système ne se contente pas d'extraire n'importe quelle

combinaison accidentelle de deux régions rectangulaires. Au contraire, il impose un certain nombre de contraintes spatiales, ce qui conduit à une augmentation certaine de la robustesse.

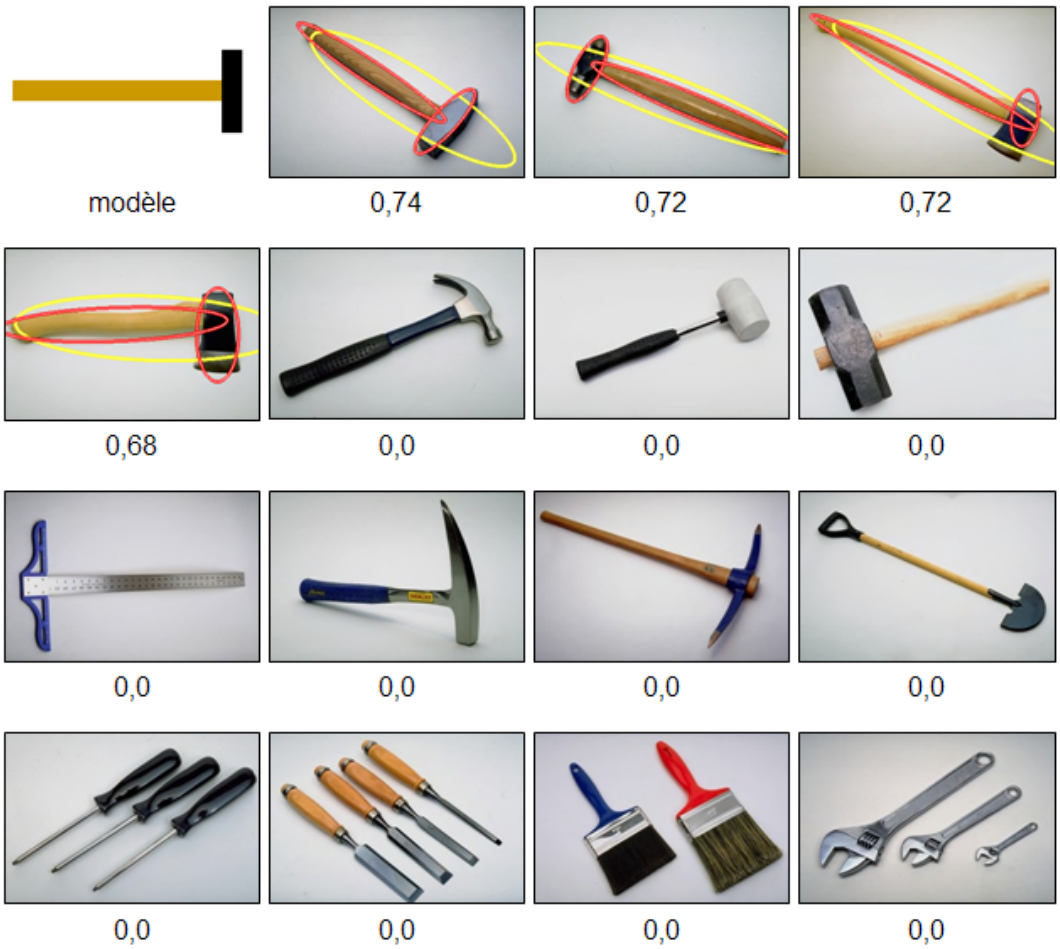


FIG. 3.16 – Exemple de résultats de la requête *marteau*.

Les deux premiers résultats (par ordre décroissant de similarité) sont effectivement des marteaux. Les deux suivants sont des haches, que l'on peut considérer comme pertinents également, du fait de leur similarité avec les marteaux. On remarque toutefois que leurs scores de similarité sont inférieurs à ceux des marteaux, du fait notamment de légères variations de forme concernant chaque sous-partie.

Une caractéristique importante de notre système est qu'il ne nécessite pas un groupement perceptuel préalable exempt d'erreur. Par exemple, le premier marteau de la figure 3.16 a obtenu un très bon score de similarité, bien que sa tête ait été fusionnée par erreur durant le groupement avec son ombre portée (voir la figure 3.4 pour les étapes principales du groupement). Néanmoins, comme la tête apparaît

tout de même plus bas dans l'arbre de régions (noeud R_7), elle peut être reconnue et appariée avec le modèle.

Deux marteaux n'ont pas été jugés similaires au modèle (c'est-à-dire que leur similarité est évaluée à 0). L'un deux (troisième image de la seconde ligne dans la figure 3.16) a été mal sur-segmenté, avant le groupement perceptuel, à cause d'une très faible variation de couleur entre le fond et la tête. La figure 3.17 montre l'image originale et l'image sur-segmentée associée. Ici, une partie du modèle (la tête) n'a jamais pu être appariée dans l'arbre de régions, puisqu'elle en était absente dès le début. Il en résulte la non-reconnaissance du modèle dans sa globalité.

(a) (b)

FIG. 3.17 – Exemple d'erreur à la sur-segmentation qui conduit à la non reconnaissance du modèle.

Concernant le deuxième marteau jugé non similaire au modèle (deuxième image de la deuxième ligne dans la figure 3.16), une moitié du manche est fusionné trop tôt avec la tête pendant le groupement perceptuel, à cause d'une forte similarité de couleur. Les étapes principales de groupement sont présentées dans la figure 3.18. Ainsi, le manche du marteau n'a jamais pu être apparié à une région de l'arbre, ce qui conduit à la non reconnaissance du modèle.

Enfin, un troisième marteau (quatrième image sur la deuxième ligne dans la figure 3.16) n'a pas non plus été jugé similaire au modèle. En effet, les descripteurs structurels de taille et position relative n'ont pas répondu suffisamment pour soutenir la similarité. Il est important de bien rappeler ici que notre système a pour but de reconnaître des structures visuelles dans les images. Ceci permet d'aller plus loin que les descripteurs bas-niveau classiques, mais le fossé sémantique empêche évidemment de pouvoir décrire un objet réel uniquement par de telles structures. Dans le cas présent, les haches sont des structures visuelles bien plus similaires au modèle que le marteau considéré. En effet, ce dernier possède une tête très volumineuse comparée au manche.

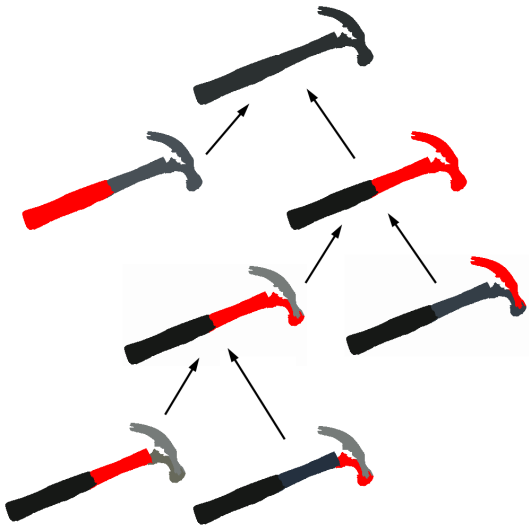


FIG. 3.18 – Exemple d'erreur dans le groupement perceptuel qui conduit à la non reconnaissance du modèle.

3.5.1.2 Modèle fortement structuré

Les descripteurs structurels deviennent particulièrement efficaces lorsque le modèle contient beaucoup de régions : l'algorithme d'appariement dispose alors de plus de critères pour trouver le bon sous-arbre optimal. Dans la figure 3.19, le système a été capable de trouver une carte *dix de pique* sous différentes positions et à différentes tailles.

Le modèle utilisé ici comporte onze régions : une de fond (représentant la carte) et dix sous objets représentant chacun des dix symboles pique. Or, du fait des conditions d'éclairage, les bords des cartes sont à peine visibles sur les images. Il en résulte que tous les fonds des cartes sont fusionnés en une seule grande région dès la sur-segmentation préalable au groupement perceptuel. Ceci rend l'appariement de la partie *fond de carte* du modèle impossible et perturbe grandement les résultats. Pour les besoins des expérimentations, dans ce paragraphe uniquement, les bords des cartes ont donc été détourés à la main afin d'éviter les erreurs de sur-segmentations. Nous verrons dans la section suivante comment s'affranchir en partie de ce problème.

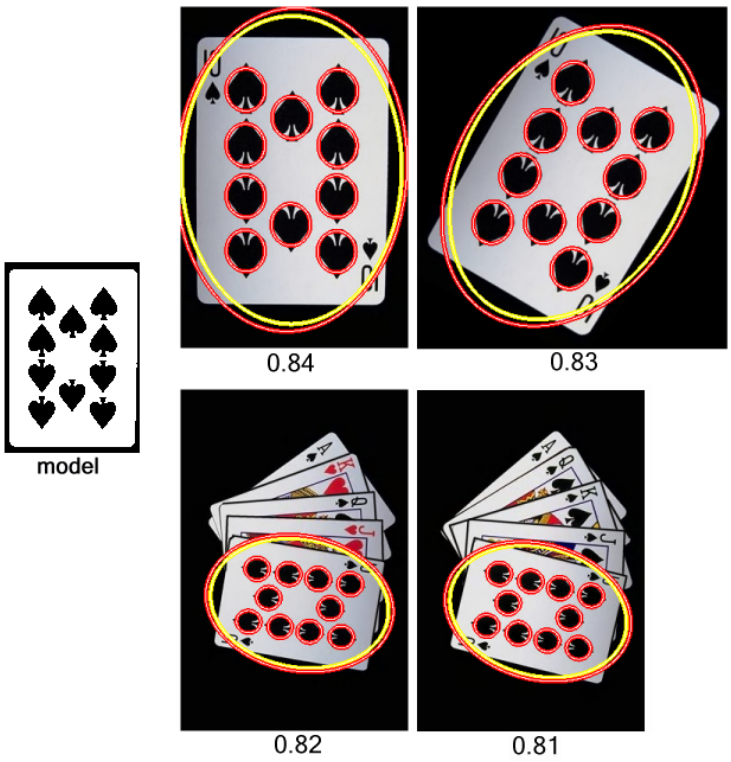


FIG. 3.19 – Exemples de résultats pour la requête *dix de pique*

Puisque le système compte les sous-objets trouvés et applique des pénalités à chaque fois qu'il en manque un, il est possible d'effectuer des opérations de classement. Par exemple, dans la figure 3.20, le système trie les cartes en fonction de leur valeur : dix de pique, puis neuf de pique, etc.

3.5.1.3 Modèle à très faible variation

Dans ce paragraphe, nous considérons un modèle dont les différentes occurences dans les images naturelles présenteront de faibles variations. En particulier, nous recherchons des drapeaux tricolores à bandes verticales. Le modèle étant très strict, les résultats sont excellents : la figure 3.21 montre les résultats sur une base de 100 images représentant toutes des drapeaux. Les six premiers résultats, par ordre décroissant de similarité, sont effectivement des drapeaux tricolores et possèdent tous la même mesure de similarité $(0,85)$. Le septième est le drapeau canadien, lui aussi tricolore, mais avec une similarité lègèrement inférieure $(0,71)$ du fait de sa bande centrale un peu plus large que les autres. La présence de la feuille d'érable est sans conséquence aucune sur la similarité.

Enfin, à ce stade, tous les drapeaux tricolores de la base ont été renvoyés et aucun n'a été manqué. Les trois derniers drapeaux renvoyés présentent également

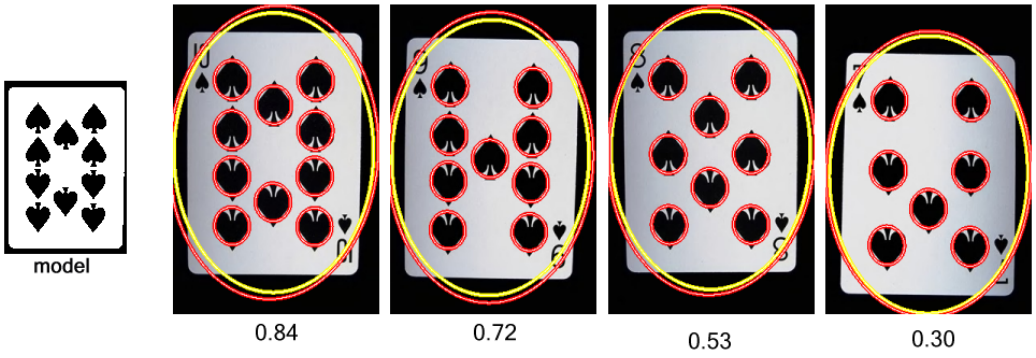


FIG. 3.20 – Exemples de classement sur des cartes, en fonction de la similarité à un modèle de référence (ici : le dix de pique).

trois zones adjacentes. Cependant, du fait de la différence d'organisation spatiale, ils obtiennent des scores bien inférieurs aux précédents, ce qui est conforme à l'intuition.

3.5.2 Résultats sur une base de 600 images

Afin de caractériser le comportement du système de manière plus approfondie, des expérimentations ont été menées sur une base de 600 images, issues de la base Corel. Cette dernière est une base largement répandue dans la communauté image, et est très fréquemment utilisée pour évaluer des systèmes d'indexation. Elle se compose de différentes classes sémantiques (par exemple : *tigre*, *aviation*), comportant chacune 100 images. Une même image peut appartenir à deux classes différentes.
Pour nos expérimentations, nous avons utilisé un échantillon de 600 images issues de cette base, et appartenant aux classes *things 1*, *things 2*, *things 3*, *tools*, *flags* et *cards*. Nous avons choisi ces classes car elles représentent toutes des objets particuliers, sur des fonds variés (relativement uniforme avec des ombres, ou complexes avec d'autres objets). Un échantillon réduit de la base est présenté en figure 3.22.

3.5.2.1 Protocole expérimental

Nous présentons maintenant le processus d'indexation - requête. Il s'appuie sur trois phases successives :

1. l'indexation de la base qui extrait pour chaque image différents descripteurs
2. l'analyse du modèle requête, qui en extrait différents descripteurs
3. la comparaison entre les descripteurs extraits du modèle et ceux issus de chaque image de la base

Durant la première phase, chaque image de la base est traitée pour en extraire ses caractéristiques. Tout d'abord, l'image est sur-segmentée par un algorithme *mean-*

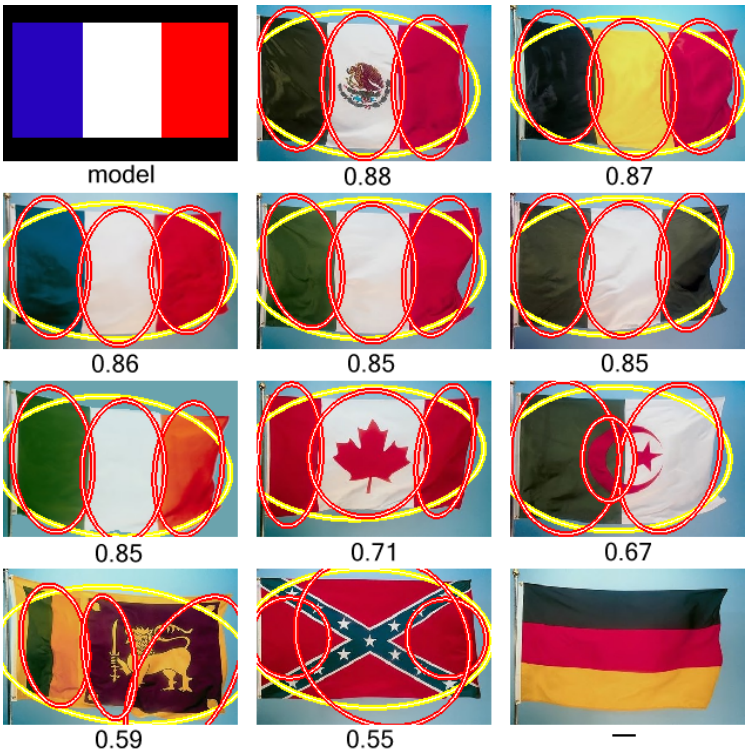


FIG. 3.21 – Résultats de la requête *drapeaux tricolores* sur une base de 100 images.

shift dont nous avons déjà parlé. Ensuite, un groupement perceptuel est effectué sur les régions issues de cette sur-segmentation. Il est mené jusqu'à n'avoir plus qu'une seule région dans l'image. De l'historique du groupement il résulte un arbre de régions, sur lequel différents descripteurs sont extraits : ART, CSS, taille, orientation, position, etc... Tous ces descripteurs sont sauvegardés dans un fichier indépendant de l'image. Cette dernière n'est alors plus requise par le système pour effectuer la requête. En outre, ce processus ne doit être conduit qu'une seule fois par image et ne prend qu'une seconde environ par image. Ce temps dépend en grande partie de la taille et de la complexité intrinsèque de l'image.

L'analyse du modèle relève exactement du même processus, mais appliqué au modèle requête : sur-segmentation, groupement perceptuel et extraction des descripteurs. Du fait de la très grande simplicité du modèle, cette étape est extrêmement rapide. Actuellement, un modèle prend la forme d'une image bitmap. On peut imaginer par la suite une interface de construction, sous la forme d'outils simple de dessins : carré, rond, ligne, point, etc...

Enfin, l'algorithme de comparaison utilise les fichiers intermédiaires de chaque image de la base pour comparer ses descripteurs extraits à ceux du modèle. Pour l'instant, aucune optimisation n'est faite et toutes les images sont comparées au

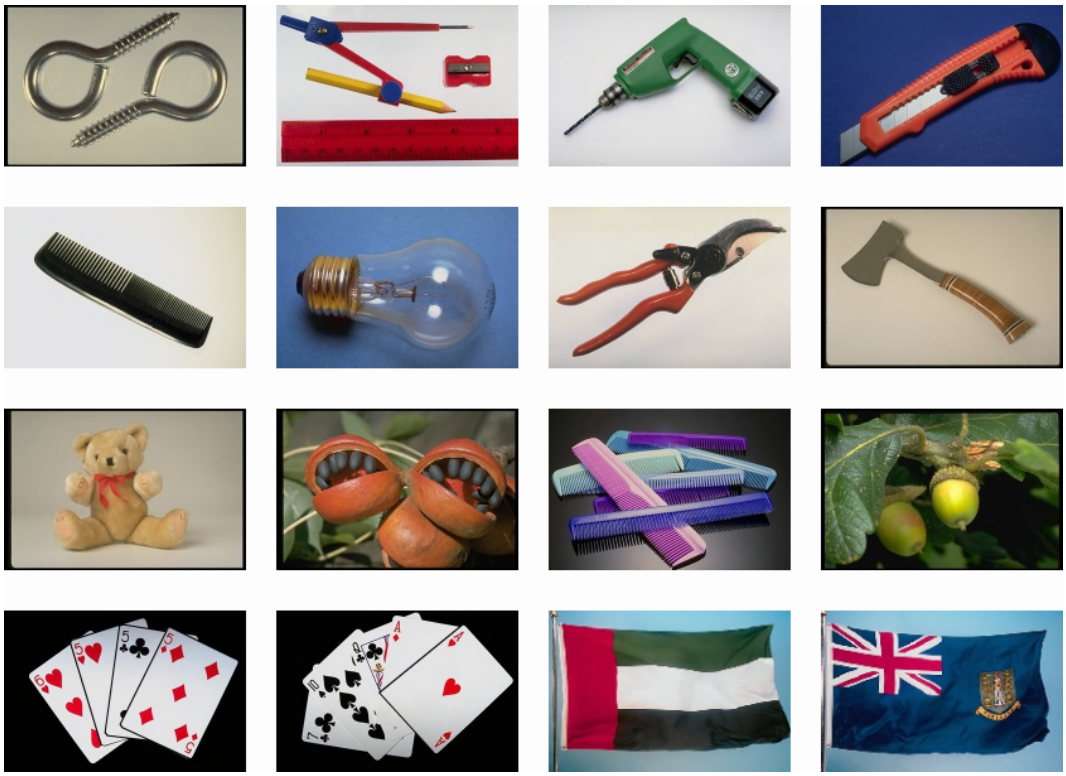


FIG. 3.22 – Exemple d'images Corel utilisées dans la base d'expérimentation.

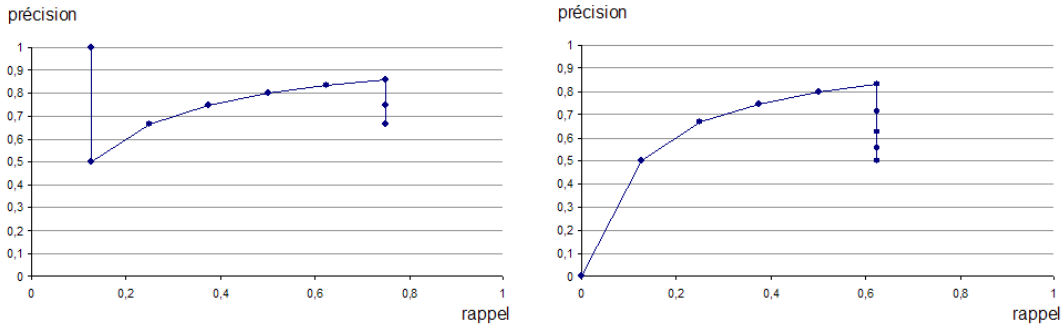
modèle. La comparaison prend donc un temps linéaire en fonction de la taille de la base. A titre indicatif, la comparaison pour une base de 600 images prend environ 5 secondes sur un Pentium 4 à 1.7 GHz. Le système reste donc très performant.

Nous disposons ainsi d'une chaîne complète de traitements pour l'indexation des images. Celle-ci est automatique, au sens où l'utilisateur n'intervient pas manuellement pour aider à la segmentation de chaque image. En outre, les temps d'indexation sont très courts.

3.5.2.2 Courbes de rappel - précision

Puisque le système renvoie une mesure quantifiée de similarité entre chaque image de la base et le modèle requête, il est possible de classer les réponses obtenues, et donc de tracer des courbes de rappel - précision.

La figure 3.23(a) présente une de ces courbes pour le modèle *marteau* composé de deux parties : le manche et la tête. Les résultats sont très satisfaisants. Plus précisément, le taux de rappel se situe aux environs de $0,75$ à son maximum (8 occurrences en tout dans la base).



(a) (b)

FIG. 3.23 – Rappel - précision pour le modèle *marteau* à 2 parties (a) et à une seule partie (b).

A titre de comparaison, une autre requête composée du même modèle mais à une seule partie a également été testée. La courbe rappel - précision correspondante est présentée dans la figure 3.23(b). On constate une perte significative tant en rappel qu'en précision. Ceci illustre l'apport de l'information structurelle lors de la requête par rapport aux approches classiques, qui considèrent un objet dans sa globalité. La figure 3.24 présente les deux modèles utilisés.

(a) (b)

FIG. 3.24 – Modèles *marteau* à 2 parties (a) et à une seule partie (b).

Lorsque les différentes occurrences du modèle dans la base présentent une très faible variation, les résultats deviennent très satisfaisants. Ainsi, la figure 3.25 montre que le rappel - précision pour le modèle *drapeau tricolore* sont excellents : toutes les occurrences sont retrouvées (6 images au total) et aucun faux positif n'apparaît.

Revenons maintenant un instant sur le modèle de carte *dix de pique*. Les résultats présentés dans la partie précédente étaient très encourageants, mais possédaient une limitation fondamentale car ils supposaient d'accentuer les contours des cartes au préalable (voir figure 3.26). En effet, le modèle utilisé comportait onze parties : une pour la carte dans sa globalité, et dix pour les différents symboles de pique. Or, puisque les bords des cartes sont à peine visibles sur les images originales, ceux-ci sont très rarement extraits. Il en résulte donc que sans correction préalable, la partie du modèle liée au fond de la carte n'est que rarement appariée dans les arbres de régions et le modèle peu reconnu.

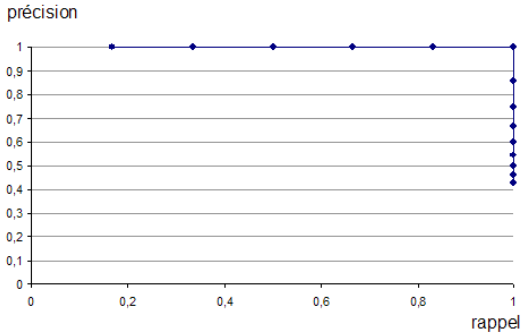
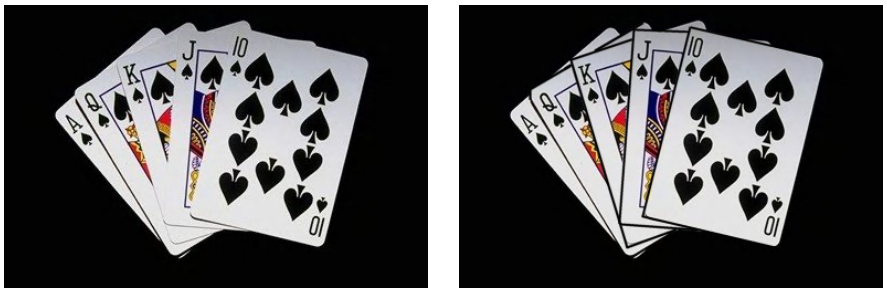


FIG. 3.25 – Rappel - précision pour le modèle *drapeau tricolore* à 3 parties.



(a) (b)

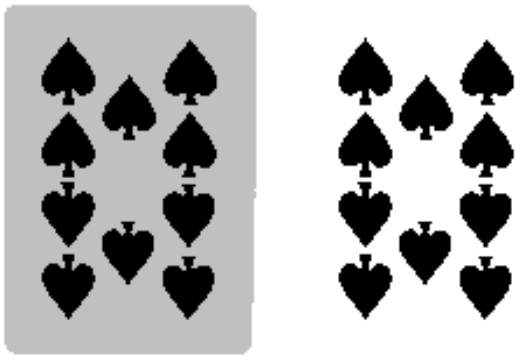
FIG. 3.26 – Exemple d'accentuation manuelle de contours (b) sur une image originale (a).

Conscients de cette limitation, nous avons expérimenté une extension, qui consiste à construire un modèle composé uniquement de parties strictement nécessaires : ici les dix symboles pique, sans le fond. Les deux variétés de modèles sont présentées dans la figure 3.27.

La courbe résultante de rappel/précision pour le modèle sans fond est présentée figure 3.28. On constate que les résultats restent très satisfaisants (forte précision) même si le rappel baisse légèrement, ce qui est dû au fait que le modèle est amputé d'une partie relativement caractéristique.

3.5.3 Autres résultats théoriques : retour sur la similarité

Il n'est pas inutile ici de revenir un instant sur la mesure de similarité proposée. Nous avons vu dans la section 1.4.2.5 de l'état de l'art, que Tversky avait mis en évidence certaines propriétés quant à la notion de jugement humain de similarité. En particulier, un résultat important concernait le fait que ce jugement ne respectait pas les axiomes définissant une distance mathématique. Or, la plupart des systèmes



(a) (b)

FIG. 3.27 – Modèles *dix de pique* à 11 parties (a) et à 10 parties (b).

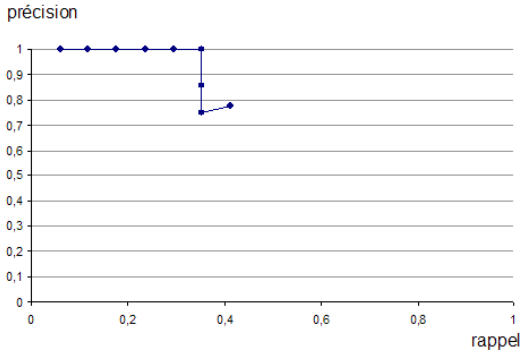


FIG. 3.28 – Rappel - précision pour le modèle *dix de pique* à dix parties.

d'indexation existants s'appuie sur cette notion de distance. Comment la fonction de similarité que nous avons proposée se positionne-t-elle face à ces questions ? C'est ce que nous nous proposons d'examiner maintenant.

3.5.3.1 Caractérisation par les axiomes de distance

Rappelons que mathématiquement, une distance d se définit comme une fonction à valeur dans $[0\,;1]$ et vérifiant les trois propriétés suivantes :

$$d(\mathbf{C}_1, \mathbf{C}_2) + d(\mathbf{C}_2, \mathbf{C}_3) \geq d(\mathbf{C}_1, \mathbf{C}_3) \text{ inégalité triangulaire} \tag{3.6}$$

$$d(\mathbf{C}_1, \mathbf{C}_2) = d(\mathbf{C}_2, \mathbf{C}_1) \text{ symétrie} \tag{3.7}$$

$$d(\mathbf{C}_1, \mathbf{C}_2) = 0 \Leftrightarrow \mathbf{C}_1 = \mathbf{C}_2 \tag{3.8}$$

On peut alors construire une mesure de similarité S à partir de d, en utilisant une fonction g strictement décroissante et en posant :

$$S = g \circ d \tag{3.9}$$

On peut alors montrer que S doit également respecter les trois axiomes d'une distance. La seule différence provient du renversement de l'inégalité triangulaire, du fait de la fonction g, décroissante.

Notons $\mathcal{S}$ la mesure de similarité que nous avons proposé. Cette dernière s'applique entre un modèle M et une image I représentée par son arbre de régions correspondant R. On note $\mathcal{S}(M, I)$. Ainsi, cette mesure de similarité est, par définition même, non symétrique. Il s'en suit que l'inégalité triangulaire ne peut pas être vérifiée par $\mathcal{S}$.

Dès lors, notre mesure de similarité n'est pas obtenue à partir d'une distance mathématique.

3.5.3.2 Caractérisation fonctionnelle

Il serait possible de chercher une expression analytique de $\mathcal{S}$, dans la mesure où une telle expression existe. Néanmoins, cela ne présenterait pas un intérêt majeur, du fait de la grande complexité de celle-ci.

Il est en revanche beaucoup plus utile de rechercher une expression fonctionnelle de $\mathcal{S}$. Sans revenir sur le détail des calculs, rappelons que cette mesure est obtenue principalement en deux temps : tout d'abord en cherchant le maximum d'appariements entre les régions R_j de l'arbre et les parties M_i du modèle, puis en calculant ensuite, grâce à ces appariements, un score global de correspondance entre le modèle M et l'arbre R.

Ainsi, $\mathcal{S}$ peut être vue comme la somme de deux influences :

- chacun des appariements entre régions R_j de l'arbre et les parties M_i du modèle tendent à augmenter la similarité
- chaque partie M_i du modèle, non appariée, tend à faire chuter la similarité

Si l'on entre un peu plus dans le détail, rappelons que chaque appariement région / partie est caractérisé par une confiance $a_{ij,k}$. Ces confiances sont combinées entre elles par le formalisme de l'évidence, qui conduit à leur renforcement mutuel. En notant Δ l'opérateur de combinaison, on peut exprimer l'influence des appariements avec :

$$\Delta_{\text{appariements}} a_{ij,k} \tag{3.10}$$

De même, chaque partie M_i du modèle, non appariée, est caractérisée par un score $b_{ij,k}$. Ces derniers sont combinés entre eux par la théorie de l'évidence, qui conduit encore une fois à leur renforcement mutuel. De la même manière, l'influence des $b_{ij,k}$ peut donc se noter par :

$$\Delta_{\text{non appariements}} b_{ij,k} \tag{3.11}$$

Enfin, toujours du fait de l'utilisation de la théorie de l'évidence pour combiner les différentes influences, une dernière combinaison va prendre en compte les deux termes précédents. Elle peut être formalisée comme suit :

$$(\Delta_{\text{appariements}} a_{ij,k}) \,.\, (1 - \Delta_{\text{non appariements}} b_{ij,k}) \tag{3.12}$$

En développant, la mesure de similarité $\mathcal{S}(M, R)$ peut alors être notée comme suit :

$$\mathcal{S}(M, I) = \Delta_{\text{appariements}} a_{ij,k} - \Delta_{\text{appariements}} a_{ij,k} . \Delta_{\text{non appariements}} b_{ij,k} \tag{3.13}$$

Il convient de noter le signe $-$ devant le deuxième terme, car il possède une influence négative sur la similarité. En effet, ce terme sert à alimenter une croyance globale en une *di-similarité*, qui se répercute a posteriori sur la similarité (voir section 3.4.2.1).

Précisons que ce dernier terme peut être interprété comme une pénalité explicite $\Delta_{\text{non appariements}} b_{ij,k}$, issue des parties du modèle non appariées, pondérée par une mesure de "doute favorable" $\Delta_{\text{appariements}} a_{ij,k}$. C'est l'utilisation du formalisme de l'évidence qui permet l'émergence de ce doute et empêche de conclure de manière abrupte à la reconnaissance ou non.

3.5.3.3 Comparaison au modèle des contrastes de Tversky

Il est intéressant de comparer notre mesure de similarité au modèle des contrastes de Tversky (1977). Rappelons que ce dernier propose une similarité $S(R, I)$ entre deux images R (requête) et I, décrites respectivement par les caractéristiques $\mathbf{C}_R$ et $\mathbf{C}_I$, telle que :

$$S(R, I) = f(\mathbf{C}_R \cap \mathbf{C}_I) - \alpha f(\mathbf{C}_R - \mathbf{C}_I) - \beta f(\mathbf{C}_I - \mathbf{C}_R) \tag{3.14}$$

$\mathbf{C}_R \cap \mathbf{C}_I$ désigne les caractéristiques communes à R et I. Plus elles sont importantes, plus la similarité augmentera.

$\mathbf{C}_R - \mathbf{C}_I$ et $\mathbf{C}_I - \mathbf{C}_R$ désignent les caractéristiques propres respectivement à R et I. Plus elles sont importantes, moins la similarité sera importante.

On constate que ce modèle est très proche du notre. En effet :

- le premier terme de Tversky est directement à mettre en relation avec notre terme $\Delta_{\text{appariements}} a_{ij,k}$. Ils ont la même signification et illustrent le fait que deux objets sont d'autant plus similaires qu'ils ont de caractéristiques communes.
- le deuxième terme de Tversky, lié aux caractéristiques propres de la requête R qui n'ont pas été retrouvées dans l'image I est, de même, à mettre en relation avec notre deuxième terme $\Delta_{\text{appariements}} a_{ij,k}.\Delta_{\text{non appariements}} b_{ij,k}$. Dans notre cas, la pénalité vient des caractéristiques du modèle M qui n'ont pas été reconnues dans l'image I.

On constate toutefois que notre mesure ne rend pas compte du dernier terme de Tversky, relatif aux caractéristiques propres de l'image I en cours d'évaluation. Il faut bien voir que notre mesure s'utilise dans le cas particulier de la comparaison d'un objet réel à un modèle. La notion de caractéristique propre à l'image en cours et non au modèle ne nous est d'aucune utilité ici.

On constate donc que notre mesure peut être vue comme une mesure de Tversky appliquée à la comparaison entre un modèle et une image. On peut toutefois noter quelques différences notables :

- Tout d'abord les travaux de Tversky s'appuyaient sur des caractéristiques à base de prédicats logiques. Notre mesure peut dès lors être vue comme une extension à des prédicats quantifiés.
- Notre mesure, comme celle de Tversky, implique un terme de pénalité sur la similarité dûe aux caractéristiques du modèle qui n'ont pas été retrouvées dans l'image. Toutefois, dans notre mesure, cette pénalité n'est pas "brute", mais est au contraire pondérée par un terme de "doute favorable" dont nous avons déjà parlé. Ceci est dû à l'utilisation de la théorie de l'évidence et permet de nuancer l'effet de la pénalité brute.

3.6 Conclusion sur l'indexation structurelle

Nous avons présenté dans cette partie un nouveau système d'indexation, dédié a à des requêtes visant à chercher des objets. Plusieurs points caractéristiques peuvent être rappelés ici.

Une requête par modèle

Nous proposons d'interroger la base en formulant des requêtes de type modèle. Ces dernières visent à conceptualiser une sorte de prototype qui synthétiserait l'objet

à rechercher. L'introduction de cette notion de modèle permet de préciser la requête, en formalisant explicitement les caractéristiques souhaitées pour chaque objet. Les paradigmes classiques, comme celui de la requête par l'exemple, ne permettent pas forcément d'inférer de la requête ce que souhaite l'utilisateur.

Vers une segmentation hiérarchique

Nous considérons que la question de la segmentation est un problème particulièrement difficile à résoudre, en particulier dans un univers non contraint comme celui de l'indexation. Nous ne cherchons donc pas à décrire chaque image par une segmentation, dont la précision ou la pertinence laisserait souvent à désirer. Au contraire, chaque image est décrite par une hiérarchie de segmentations (arbre de régions), à différents niveaux de détail. Nous avons vu que ceci permet de s'affranchir d'un certain nombre d'erreurs de segmentation lors de la requête.

Intégration de données structurelles

La comparaison de la requête à un arbre de régions se fait en deux temps. Tout d'abord le système cherche à apparier le maximum de parties à différentes régions de l'arbre, sur la base de plusieurs critères : formes des parties, mais aussi organisation spatiale des parties par rapport à l'objet entier. Ensuite, ces différents appariements servent au calcul d'une similarité globale entre le modèle et l'arbre.
Les différents descripteurs utilisés pour ces appariements sont en outre combinés avec le formalisme de la théorie de l'évidence, ce qui permet une modélisation fine des différentes interactions souhaitées. Il en résulte une mesure de similarité assez robuste dans la mesure où elle nécessite l'activation de plusieurs descripteurs pour son incrémentation.

Des résultats efficaces et prometteurs

Différents tests ont été menés, notamment sur une base de 600 images naturelles, afin de montrer le bon comportement de notre système. A ce stade, il convient de remarquer que ce dernier est limité à des requêtes impliquant des objets aux parties bien définies (en tant que régions). Ainsi, il serait extrêmement complexe d'utiliser notre système pour formuler une requête de recherche de visage par exemple. En effet, les différents composants d'un visage ne peuvent être défini par des régions aux contours nettement dessinés.
Toutefois, il demeure possible de traiter des requêtes dans lesquelles seulement une partie des objets ont des frontières bien définies. Ainsi, nous avons vu que pour la requête *dix de pique*, le fond de la carte posait problème puisqu'il était souvent absent des segmentations. Une solution a consisté à simplement omettre cette partie dans le modèle. Ce faisant, les résultats restent honorables.

La limitation de ce genre de solution est évidente : comment un utilisateur peut-il savoir à l'avance quelle partie intégrer dans son modèle et lesquelles éliminer. Ceci pose la délicate question de la conception d'interface pour la création et la modification de requête à des systèmes d'indexation.

4

Conclusion et perspectives

Deux éléments principaux de la chaine de traitement de l'indexation ont été considérés pendant cette thèse : l'étape cruciale de segmentation d'une part et la combinaison des différents descripteurs lors la requête d'autre part. Il faut toutefois garder à l'esprit qu'un prototype complet a été developpé, et qu'il permet de prendre en charge la totalité des phases d'indexation et de requête à une base d'images (rappel de la chaîne en figure 4.1). En outre, le système d'indexation est automatique, au sens où, par exemple, il ne requiert pas une assistance de l'utilisateur lors de la segmentation de chaque image. Les performances sont très satisfaisantes d'après les premières évaluations de rappel et de précision dont nous disposons. Enfin, les temps de traitements de cette chaîne sont excellents puisqu'une base de 600 images est indexée en quelques minutes, de manière définitive, et qu'une requête sur cette base ne prend que quelques secondes.

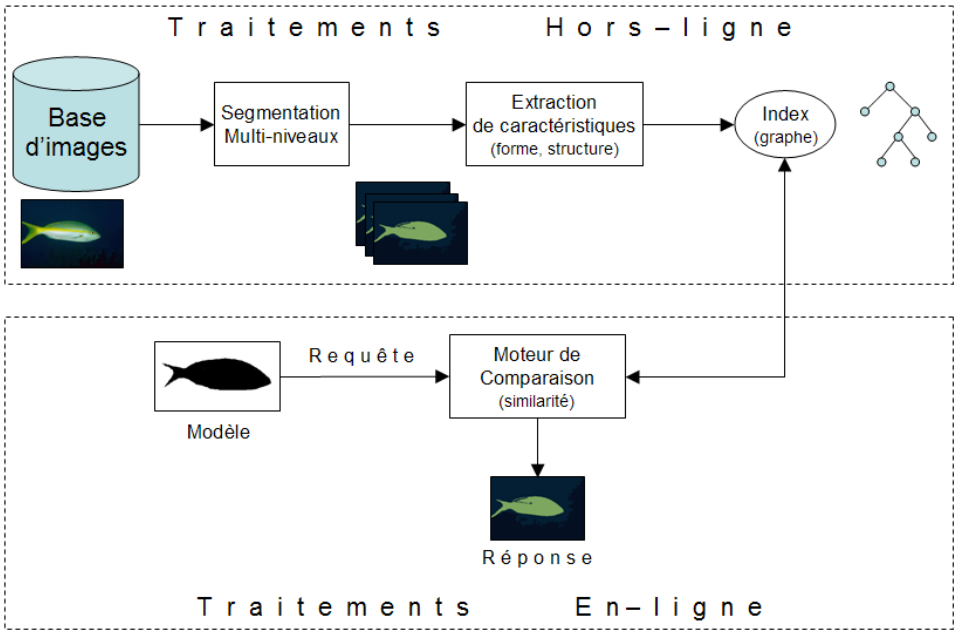


FIG. 4.1 – Chaine de traitements pour l'indexation, proposée dans cette thèse.

Il convient désormais de résumer nos principales contributions et de proposer quelques perspectives, pour chacun des deux points principaux abordés dans cette thèse.

Segmentation d'image et groupement perceptuel

Contributions

Nous avons proposé un système de segmentation hiérarchique, basé sur des critères perceptuels. Classiquement, les mécanismes de segmentations partitionnent une

image en différentes régions suivant des propriétés locales de couleur ou de texture. Ils restent donc fondamentalement limités pour extraire de l'image des zones qui correspondent à des objets réels. En effet, ces derniers comportent souvent de nombreux artefacts de couleurs dûs, par exemple, à des variations dans les conditions d'éclairage. En outre, un objet réel est très souvent composé de différentes parties, chacune d'entre elles possédant des descripteurs très différents de couleur ou de texture. Ces différentes parties ne peuvent donc pas être considérées comme une seule entité sur la seule base de descripteurs bas-niveaux.
Ainsi, en utilisant d'autres critères, il est possible de rafiner la qualité de la segmentation. Citons par exemple l'utilisation de propriétés géométriques entre différentes régions adjacentes comme la continuité ou le parallélisme. Plus généralement, les critères que nous avons utilisés sont inspirés des résultats d'études psycho-visuelles sur la perception humaine.

Si de nombreux travaux ont été publiés sur cette question, très peu ont tenté de manipuler à la fois des primitives régions et contours. Or, le fait de considérer conjointement ces deux primitives permet un gain certain de robustesse lors d'une segmentation préalable. En outre, ceci permet, lors de la description de chaque propriété du groupement perceptuel, d'utiliser la primitive la plus adaptée.
Une deuxième contribution importante sur ce travail concerne la question de la combinaison ou de l'interaction des différentes propriétés de groupement perceptuel entre elles. Nous avons proposé un mécanisme d'interaction collaborative, modélisé par la théorie de l'évidence, qui impose à plusieurs propriétés d'être activées simultanément afin de déclencher un groupement entre deux régions. Encore une fois, ceci permet un gain certain de robustesse, puisqu'une propriété ne peut, à elle seule, faire fusionner deux régions.

Perspectives

Une perspective intéressante relative à ce travail concerne l'extraction des différentes hypothèses de groupements envisageables. Pour l'instant, ces dernières sont issues du graphe d'adjacence des régions. Ainsi, chaque paire de voisins immédiats dans le graphe donne naissance à une hypothèse de groupement. Or, il serait évidemment utile de considérer une vision plus élargie des groupements potentiels. Par exemple, on pourrait considérer pour chaque noeud, non plus uniquement ses voisins immédiats, mais tous ceux d'un ordre supérieur (deux, trois, voire plus). Ceci poserait toutefois la question de la limite de voisinage à considérer. En outre, cette extension poserait problème pour la description de certaines propriétés : comment décrire la compacité de deux régions non adjacentes ?

Une deuxième perspective concernant ce travail a trait au mécanisme de réduction proprement dit du graphe d'adjacence, une fois que les hypothèses sont valuées d'une confiance globale. Pour l'instant, nous utilisons un algorithme glouton qui nous assure de sa convergence, mais qui peut évidemment être piégé par des minima

locaux. L'utilisation d'algorithme de graphes, comme celui de la coupe minimum, éventuellement optimisé par différentes heuristiques, pourrait permettre de corriger ce problème et permettrait vraisemblablement de gagner en qualité de résultats.

Combinaison de descripteurs lors de la requête

Contributions

Nous ne prétendons pas avoir proposé une méthode de segmentation robuste dans un univers non contraint. Ceci est en effet une question extrêmement délicate. Au contraire, nous assumons le fait que notre algorithme de segmentation commet des erreurs, comme tous les autres processus uniquement basés sur le signal "image" et sans connaissance externe de contexte ou autre. Une de nos contributions concerne alors un mécanisme d'indexation qui permet de prendre en compte certaines de ces erreurs issues de la segmentation. Le principe consiste à ne pas décrire une image par une segmentation unique, qui comporterait fatalement des imprécisions et qui, surtout, serait fixée à un niveau de détail donné. Au contraire, nous décrivons chaque image par une hiérarchie de groupements issus de la segmentation. Le fait d'utiliser cette hiérarchie permettra de corriger certaines erreurs dûes à la segmentation, mais aussi d'utiliser des descripteurs structurels, et surtout de considérer chaque image à des niveaux de détail différents, en fonction de la requête proposée.

Lors de la comparaison d'une image à un modèle, nous proposons une modélisation fine de la combinaison des différents descripteurs (forme, structure), toujours avec la théorie de l'évidence. Cette dernière se révèle être une extension d'un modèle théorique bien connu, celui de Tversky.

Un système complet de traitement a été proposé (THIS ou *THings-oriented Image retrieval System*), depuis la segmentation, jusqu'à la requête. Les résultats ont été testés sur une base de 600 images naturelles. Même si ce nombre reste faible en regard des tailles des bases d'images existantes, il reste assez important en terme d'évaluation d'un travail de recherche. Les résultats sont très satisfaisants d'après les premières évaluations de rappel et de précision dont nous disposons. En outre, le système est très efficace en terme de temps de calcul, tant pour l'indexation (quelques minutes pour la base entière) que lors de la requête (quelques secondes, sans optimisation).

Perspectives

Une perspective future de recherche consisterait tout d'abord à être capable de pouvoir corriger encore plus d'erreurs issues de la segmentation lors de la requête. En effet, si une partie d'un objet a été, très tôt dans la hiérarchie de groupement, séparée en deux régions, sans jamais que ces dernières soient fusionnées par la suite,

notre système n'est pas, en l'état, capable de retrouver la partie de modèle correspondant. Ainsi, il serait utile de pouvoir tout de même apparier les régions fragmentées, chacune à un *morceau* de parties de modèle. Pour ceci, il serait envisageable, une fois le modèle soumis, de fragmenter chacune de ses parties arbitrairement, et d'essayer ensuite d'apparier les régions des hiérarchies de groupements à chacun des fragments.

En outre, un aspect important relatif à la reconnaissance d'objets n'a pas été abordé dans ce manuscrit. Il concerne la capacité du système à retrouver un même objet sous différents points de vue. Ceci serait possible à mettre en oeuvre dans notre système, en ajoutant une étape de traitements. Ainsi, chaque modèle utilisé serait en fait une modélisation d'un point de vue de l'objet. Pour retrouver l'objet sous différents points de vue, il conviendrait alors de faire, par exemple, trois modèles : un de face, un de profil et un de trois quarts. Chaque modèle serait traité séparemment, comme présenté dans la partie précédente. Puis, une dernière étape combinerait, toujours au sens de Shafer, les résultats obtenus sur chacun des points de vue.

Par ailleurs, une piste importante d'évolution de notre système, tant du point de vue théorique que pratique, concerne la pondération des descripteurs lors de la comparaison. En effet, même si les descripteurs sont combinés par la théorie de l'évidence, ils peuvent toujours être pondérés. Ceci est même primordial, puisque certaines requêtes souhaiteront insister sur un certain type de descripteurs, alors que d'autres non. Par exemple, la requête *marteau* se doit d'insister sur la bonne correspondance des descripteurs structurels, plutôt que sur les descripteurs de forme pour chacune des parties. En effet, il n'existe pas, par exemple, une seule forme pour une tête de marteau. En revanche, l'agencement structurel des différentes parties est relativement fixe.
Une première évolution serait donc d'intégrer un retour utilisateur dans le processus de requête. Ainsi, le système présenterait une série de résultats "bruts" à l'utilisateur, qui préciserait, pour chacun, s'il est satisfait ou non de la réponse. Ainsi, le système disposerait à ce stade d'une information supplémentaire pour éventuellement corriger la pondération de certains descripteurs par rapport à d'autres. Ceci pourrait permettre d'affiner les résultats.

Toujours dans cette optique de pondération des différents descripteurs, il est parfaitement envisageable de construire des bases d'apprentissage contenant des images représentatives d'une classe donnée d'objets. Les différentes images de chaque catégorie pourraient alors servir à un apprentissage supervisé via l'initialisation d'une jeu de paramètres pour chaque classe d'objets. On disposerait alors de différents systèmes capables de reconnaître directement une classe d'objets, sans aucun autre paramétrage.

Il est important de noter que toutes ces questions sur la pondération relative des descripteurs renvoient en fait à la notion plus large de similarité entre objets. Notre communauté de "traiteurs d'images" adresse cette question en considérant la partie signal de l'image. Nous avons vu que l'introduction de descripteurs de

formes ou de structures, par exemple, permet d'affiner les indexs, et de réduire une partie du fossé sémantique. Néanmoins, la notion de similarité entre objets renvoie aussi à des considérations sémiotiques. En particulier, une dimension fondamentale est liée à l'utilisateur en tant que tel : recherche-t-il, pour une requête donnée, un objet répondant à la fonctionalité du marteau, ou bien un objet dont la structure, la forme, s'apparente à celle d'un marteau ? Les deux cas sont évidemment bien différents. Ainsi, il s'agit de réussir à saisir ce que cherche réellement l'utilisateur, au-delà du modèle et du signal.

Bibliographie

Ackermann, A., Massmann, A., Posh, S., Sagerer, G. et Schlüter, D. « Perceptual grouping of contour segments using markov random fields ». *Pattern Recognition and Image Analysis*, 7(1) :11–17. **1997**. pages 56

Berretti, S., Del Bimbo, A. et Pala, P. « Retrieval by Shape Similarity with Perceptual Distance and Effective Indexing ». *IEEE Transactions on Multimedia*, 2(4) :225–239. **2000**. pages 25

Biederman, I. « Recognition by Components : A Theory of Human Image Understanding ». *Psychological Review*, 94 :115–147. **1987**. pages 25

Carson, C., Belongie, S., Greenspan, H. et Malik, J. « Blobworld : image segmentation using expectation-maximization and its application to image querying ». *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(8) :1026–1038. **2002**. pages 20, 29, 49

Chang, S. et Hsu, A. « Image Information Systems - Where do we go from here ». *IEEE Transactions on Knowledge and Data Engineering*, 4(5) :431–442. **1992**. pages 31

Chong, C.-W., Raveendran, P. et Mukundan, R. « An Efficient Algorithm for Fast Computation of Pseudo-Zernike Moments. ». *International Journal of Pattern Recognition and Artificial Intelligence*, 17(6) :1011–1023. **2003**. pages 25

Christoudias, C. M., Georgescu, B. et Meer, P. « Synergism in Low Level Vision ». Dans *IEEE International Conference on Computer Vision (ICCV'02)*, volume IV, pages 150–155. IEEE Computer Society. **2002**. pages 54

Comaniciu, D. et Meer, P. « Robust analysis of feature spaces : color image segmentation ». Dans *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'97)*, pages 750–757. **1997**. pages 49, 53, 54, 69

Del Bimbo, A. et Pala, P. « Visual Image Retrieval by Elastic Matching of User Sketches ». *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(2) :121–132. **1997**. pages 25

DEMPSTER, A., LAIRD, N. et RUBIN, N. « Maximum Likelihood from Incomplete Data via the EM Algorithm ». *Journal of Royal Statistical Society, Ser. B*, 39(1) :1–38. **1977**. pages 29

DESOLNEUX, A. et MOREL, J. « A grouping principle and four applications ». *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(4) :508–513. **2003**. pages 55, 56

« Dublin Core Metadata Element Set Version 1.1 : Reference Description ». Dublin Core Metadata Initiative (DCMI). **2003**. *http ://dublincore.org/documents/2003/06/02/dces/, (consulté le 21/06/06)*. pages 9

FERGUS, R., PERONA, P. et ZISSERMAN, A. « Object Class Recognition by Unsupervised Scale-Invariant Learning ». Dans *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'03)*, pages 264–271. **2003**. pages 34

FLICKNER, M., SAWNEY, H., NIBLACK, W., ASHLEY, J., HUANG, Q., DOM, B., GORKANI, M., HAFNER, J., LEE, D. et PETKOVIC, D. « Query by Image and Video Content : the QBIC System ». *IEEE Special Issue on Content-Based Picture Retrieval System*, 28(9) :23–32. **1995**. pages 23

FORSYTH, D. et FLECK, M. « Automatic Detection of Human Nudes ». *International Journal of Computer Vision*, 32(1) :63–77. **1999**. pages x, 31, 32, 52

FORSYTH, D., MALIK, J. et WILENSKY, R. « Searching for Digital Pictures ». *Scientific American*, 276(6) :72–77. **1997**. pages 11, 93

GONZALEZ, R. C. et WOODS, R. E. *Digital Image Processing*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA. **2001**. pages 25

HAFNER, J., SAWHNEY, H. S., EQUITZ, W., FLICKNER, M. et NIBLACK, W. « Efficient Color Histogram Indexing for Quadratic Form Distance Functions ». *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(7) :729–736. **1995**. pages 40

HANSEN, T. *A neural model of early vision : contrast, contours, corners and surface.*. Thèse de doctorat, University of ULM, Germany. **2002**. pages x, 55

HARALICK, R. et SHAPIRO, L. *Computer and robot vision*. Addison Wealey Publishing, USA. **1992**. pages 15

HARRIS, C. et STEPHENS, M. « A combined corner and edge detector ». Dans *4th Alvey Vision Conference*, pages 147–151. **1988**. pages x, 29, 30

HU, M. « Visual pattern recognition by moment invariants ». *IRE Transactions on Information Theory*, IT(8) :179–187. **1962**. pages 25

Huang, J., Kumar, S., Mitra, M., Zhu, W. et Zabih, R. « Spatial Color Indexing and Applications ». *International Journal of Computer Vision*, 35(3) :245–268. **1999**. pages 22

Idrissi, K., Lavoue, G., Ricard, J. et Baskurt, A. « Object of interest-based visual navigation, retrieval, and semantic content identification system ». *Computer Vision and Image Understanding*, 94(1-3) :271–294. **2004**. pages 23, 57, 85, 86

Iivarinen, J. et Visa, A. « Shape recognition of irregular objects ». Dans D. Casasent, éditeur, *Intelligent Robots and Computer Vision XV : Algorithms, Techniques, Active Vision, and Materials Handling (SPIE '96)*, pages 25–32. **1996**. pages 25

Irani, P. et Ware, C. « Diagramming information structures using 3D perceptual primitives ». *ACM Transactions on Computer-Human Interaction (TOCHI)*, 10 :1–19. **2003**. pages 25

Itti, L., Koch, C. et Niebur, E. « Model of Saliency-Based visual attention for rapid scene analysis ». *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11) :1254–1259. **1998**. pages x, 31, 47, 58, 59, 60

Jeannin, S. « MPEG-7 Visual part of experimentation model, Version 9.0 ». Dans *ISO/IEC JTC1/SC29/WG11/N3914, 55th Mpeg Meeting*. Pisa, Italy. **2001**. pages 27, 39, 41

Jorgensen, C. *Image Retrieval : Theory and Research*. Scarecrow Press, Lanham, USA, 340 p. **2003**. pages 8

Kang, H. et Walker, E. « Multilevel grouping : combining bottom-up and top-down reasoning for object recognition ». Dans *International Conference on Pattern Recognition (ICPR'94)*, pages 559–562. **1994**. pages 56

Kanizsa, G. *Organization in vision*. Praeger, New-York, USA, 267 p. **1979**. pages 50, 68

Khotanzad, A. et Hong, Y. H. « Invariant Image Recognition by Zernike Moments ». *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(5) :489–497. **1990**. pages 25

Kim, W.-Y. et Kim, Y.-S. « A New region-Based Shape Descriptor ». Dans *Mpeg Meeting, TR 15-01*. Pisa. **1999**. pages 25, 26, 39, 100

Koch, C. et Ullman, S. « Shifts in selection in visual attention : towards the underlying neural circuitry ». *Human Neurobiology*, 4(4) :219–227. **1985**. pages 30

KOFFKA, K. *Principles of Gestalt Psychology*. Harcourt, New-York, 720 p. **1935**. pages 50, 68, 70

LAFON, Y. et BOS, B., « Describing and Retrieving Photographs Using RDF and http ». World Wide Web Consortium (W3C). **2000**. *http ://www.w3.org/TR/2000/NOTE-photo-rdf-20000928*. pages 10

LE MEUR, O., LE CALLET, P., BARBA, D. et THOREAU, D. « Performance assessment of a visual attention system entirely based on a human vision modeling ». Dans *International Conference on Image Processing (ICIP)*, volume 4. **2004**. pages x, 60

LEVY-SCHOEN, A. « Exploration et connaissance de l'espace visuel sans vision périphérique ; quelques données sur le comportement oculomoteur de l'adulte normal ». *Journal Psychologique*, 39(1) :77–91. **1976**. pages 37

LOWE, D. *Perceptual Organization and Visual Recognition*. Kluwer, Boston, 162 p. **1985**. pages 49, 51, 54, 55

LOWE, D. « Distinctive Image Features from Scale-Invariant Keypoints ». *International Journal of Computer Vision*, 60(2) :91–110. **2004**. pages 36

LUO, J. et GUO, C. « Perceptual grouping of segmented regions in color images ». *Pattern Recognition*, 36(12) :2781–2792. **2003**. pages 57, 86

LYMAN, P. et VARIAN, H., « How much information ». University of Berkeley. **2003**. *http ://www.sims.berkeley.edu :8000/research/projects/how-much-info-2003/, (consulté le 21/06/06)*. pages 1

MALLAT, S. *A Wavelet Tour of Signal Processing*. Academic Press, London ; 2nd edition, 637 p. **1999**. pages 18

MANJUNATH, B. S. et MA, W. « Texture features for browsing and retrieval of image data ». *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI - Special issue on Digital Libraries)*, 18(8) :837–42. **1996**. pages 18

MARCELJA, S. « Mathematical description of the response of simple cortical cells ». *Journal of Optical Society of America*, A 70(11) :1297–1300. **1980**. pages 18

MARR, D. *Vision*. Freeman, San Francisco, 397 p. **1982**. pages 53

MARTIN, D., FOWLKES, C., TAL, D. et MALIK, J. « A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics ». Dans *IEEE International Conference of Computer Vision (ICCV)*, volume 2, pages 416–423. **2001**. pages 79

MEHROTRA, R. et GARY, J. E. « Similar-Shape Retrieval In Shape Data Management ». *Computer*, 28(9) :57–62. **1995**. pages 25

MILANESE, R., BOST, J. et PUN., T. « A bottom-up attention system for active vision ». Dans *10th European Conference on Artificial Intelligence (ECAI-92)*, pages 808–810. **1992**. pages 60

MOHAN, R. et NEVATIA, R. « Perceptual Organization for Scene Segmentation and Description ». *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(6) :616–635. **1992**. pages 56

MOKHTARIAN, F. et MACKWORTH, A. « A theory of multiscale, curvature-based shape representation for planar curve ». *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14 :789–805. **1992**. pages ix, 25, 27, 28, 100

MURINO, V., REGAZZONI, C. et FORESTI, G. « Grouping as a searching process for minimum-energy configurations of labelled random fields ». *Computer Vision and Image Understanding*, 64(1) :157–174. **1996**. pages x, 56, 58

OLIVA, A. « Gist of a scene ». *Neurobiology of Attention*, Itti L., Rees G., Tsotsos J.K Elsevier Academic Press, Amsterdam :251–256. **2005**. pages 37, 53

OLIVA, A., TORRALBA, A., CASTELHANO, M. S. et HENDERSON, J. M. « Top-down control of visual attention in object detection ». Dans *IEEE Inernational Conference on Image Processing*. **2003**. pages 55, 72

OSADA, R., FUNKHOUSER, T., CHAZELLE, B. et DOBKIN, D. « Matching 3D Models with Shape Distributions ». Dans *International Conference on Shape Modeling & Applications (SMI-01)*, page 154. IEEE Computer Society, Washington, DC, USA. **2001**. pages 43

PAVLIDIS, T. et HOROWITZ, S. « Segmentation of plane curves ». *IEEE Transaction on Computers*, 23(8) :860–870. **1974**. pages 70

PETERSON, T. *Introduction to the Art and Architecture Thesaurus*. Oxford University Press, New-York, 250 p. **1994**. pages 9

PETRAKIS, E. et FALOUTSOS, C. « Similarity searching in medical image databases ». *IEEE Transactions on Knwoledge and Data Engineering*, 9(3) :435–447. **1997**. pages 31

PRIÉ, Y., MILLE, A. et PINON, J. « A Context-Based Audiovisual Representation Model for Audiovisual Information Systems ». Dans *Context'99, Second International and Interdisciplinary Conference on Modeling and using Context*, volume 1688, pages 296–309. **1999**. pages 10

ROS, J., LAURENT, C., JOLION, J. et SIMAND, I. « Comparing String Representations and Distances in a Natural Images Classification Task ». Dans *5th IAPR-TC-15 workshop on graph-based representations*, pages 72–81. **2005**. pages 30

ROUSSEY, C. *Une méthode d'indexation sémantique adaptée aux corpus multilingues*. Thèse de doctorat, Insa, Lyon, France, 197 p. **2001**. pages 9

SANTINI, S. *Exploratory image database, content-based retrieval*. Academic Press, San Diego, 613 p. **2001**. pages 7, 41

SANTINI, S. et JAIN, R. « Similarity is a Geometer ». *Multimedia Tools and Applications*, 5(3) :277–306. **1997**. pages 43

SARKAR, S. et BOYER, K. « Integration, Inference, and Management of Spatial Information Using Bayesian Networks : Perceptual Organization ». *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(3) :256–274. **1993**a. pages 52

SARKAR, S. et BOYER, K. « Perceptual organization in computer vision : a review and a proposal for a classificatory structure ». *IEEE Transactions on Systems, Man, and Cybernetics*, 23(2) :388–399. **1993**b. pages 51, 52

SARKAR, S. et BOYER, K. « A computational structure for preattentive perceptual organization : graphical enumeration and voting method ». *IEEE Transactions on Systems, Man, and Cybernetics*, 24(2) :246–267. **1994**. pages x, 56, 57

SCHNEIDERMAN, H. et KANADE, T. « A Statistical Method for 3D Object Detection Applied to Faces and Cars ». Dans *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'00)*. **2000**. pages x, 32, 34, 35

SCLAROFF, S. et LIU, L. « Deformable shape detection and description via model-based region grouping ». *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(5) :475–489. **2001**. pages 52

SEKITA, I., KURITA, T. et OTSU, N. « Complex Autoregressive Model for Shape Recognition ». *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(4) :489–496. **1992**. pages 25

SHAFER, G. *A Mathematical Theory of Evidence*. Princeton University Press, Princeton, 297 p. **1976**. pages 58, 63, 102, 141

SMEULDERS, A., WORRING, M., SANTINI, S., GUPTA, A. et JAIN, R. « Content-Based Image Retrieval at the End of the Early Years ». *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12) :1349–1380. **2000**. pages 7, 13, 20, 21

SMITH, R. et CHANG, S. « Integrated Spatial and Feature Image Query ». *Multimedia Systems*, 7(2) :129–140. **1999**. pages 31

SONKA, M., HLAVAC, V. et BOYLE, R. *Image Processing : Analysis and Machine Vision*. PWC Pub., Pacific Grove (CA), 770 p. **1999**. pages 25

STRICKER, M. et DIMAI, A. « Color Indexing with Weak Spatial Constraints ». Dans *Storage and Retrieval for Image and Video Databases (SPIE Š96)*, volume IV 2670, pages 29–40. **1996**. pages 40

SWAIN, M. et BALLARD, D. « Color indexing ». *International Journal of Computer Vision*, 7(1) :11–32. **1991**. pages 22, 39

TEAGUE, M. « Image analysis via the general theory of moments ». *Journal of the Optical Society of America*, 70(8) :920–930. **1980**. pages 25

TIENG, Q. M. et BOLES, W. W. « Recognition of 2D Object Contours Using the Wavelet Transform Zero-Crossing Representation ». *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(8) :910–916. **1997**. pages 25

TORRALBA, A. et A.OLIVA. « Semantic organization of scenes using discriminant structural templates ». Dans *IEEE International Conference on Computer Vision (ICCV)*, volume 2, pages 1253–1258. **1999**. pages 18

TORRALBA, A., MURPHY, K. et FREEMAN, W. « Sharing Features : Efficient Boosting Procedures for Multiclass Object Detection ». Dans *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 762–769. **2004**. pages 36

TUYTELAARS, T. et GOOL, L. J. V. « Content-Based Image Retrieval Based on Local Affinely Invariant Regions ». Dans *Visual Information and Information Systems*, pages 493–500. **1999**. pages 29

TVERSKY, A. « Features of similarity ». *Psychological Review*, 84(4) :327–352. **1977**. pages x, 41, 42, 94, 121

TVERSKY, A. et GATI, I. « Similarity, separability, and the triangle inequality ». *Psychological Review*, 89 :123–154. **1982**. pages 41

VAILAYA, A., JAIN, A. et ZHANG, H. « On Image Classification : City Images vs. Landscapes ». *Pattern Recognition*, 31(12) :1921–1935. **1998**. pages 32, 34

VAN DERWAAL, H. *ICONCLASS : An inconographic classification system*. Rapport technique, Technical report, Royal Dutch Academy of Sciences (KNAW), Netherlands. **1985**. pages 9

VAN OTTERLOO, P. J. *A contour-oriented approach to shape analysis*. Prentice Hall International (UK) Ltd., Hertfordshire, UK, 368 p. **1991**. pages 25

VASSEUR, P., PEGARD, C., MOUADDIB, E. et DELAHOCHE, L. « Perceptual organization approach based on Dempster-Shafer theory ». *Pattern Recognition*, 32(8) :1449–1462. **1999**. pages 58

« RDF / XML Syntax Specification, W3C Recommendation ». World Wide Web Consortium (W3C). **2004**. *http ://www.w3.org/TR/2004/REC-rdf-syntax-grammar-20040210/, (consulté le 21/06/06)*. pages 10

Wang, J., Li, J. et Wiederhold, G. « SIMPLIcity : Semantics-Sensitive Integrated Matching for Picture Libraries ». *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(9) :947–963. **2001**. pages 20, 29, 49

Zaharia, T. et Prêteux, F. « Comparative study for 3D and 2D/3D Shape Descriptors ». *Research Report ISO/IEC JTC1/SC29/WG11, MPEG04/10657, Munchen, Germany*. **2004**. pages 24, 26, 100

Zhang, D. et Lu, G. « A Comparison of Shape Retrieval Using Fourier Descriptors and Short-Time Fourier Descriptors ». Dans *IEEE Pacific Rim Conference on Multimedia (PCM '01)*, pages 855–860. **2001**. pages 25

Zhang, D. et Lu, G. « Generic Fourier descriptor for shape-based image retrieval ». Dans *IEEE International Conference on Multimedia and Expo (ICME '02)*, volume 1, page 425Ű428. Lausanne, Switzerland. **2002**. pages 25

Zhang, D. et Lu, G. « Review of shape representation and description techniques ». *Pattern Recognition*, 37 :1–19. **2004**. pages ix, 24, 100

A

Publications de l'auteur

Revues internationales avec comité de lecture

1. Zlatoff, N., Tellez, B. et Baskurt, A. « Combining local belief from low-level primitives for perceptual grouping ». *Pattern Recognition, en deuxième révision*.

Conférences internationales avec comité de lecture

1. Zlatoff, N., Ryder, G., Tellez, B. et Baskurt, A. « Content-Based Image Retrieval : on the Way to Object FeaturesT̃, ». Dans *International Conference on Pattern Recognition, ICPR 2006, à paraitre*.
2. Zlatoff, N., Tellez, B. et Baskurt, A. « Region-based perceptual grouping : a cooperative approach based on Dempster-Shafer theory ». Dans *IS&T / SPIE Symposium on Electronic Imaging, Image Processing : Algorithms and Systems V*, pages 244–254. San José, USA. **2006**.
3. Zlatoff, N., Tellez, B. et Baskurt, A. « Image Understanding and Scene Models : a Generic Framework Integrating Domain Knowledge and Gestalt Theory, ». Dans *IEEE International Conference on Image Processing, ICIP 2004*, pages 2355–2358. Singapore. **2004**.

Conférences nationales avec comité de lecture

1. Zlatoff, N., Tellez, B. et Baskurt, A. « Groupement perceptuel pour la reconnaissance d'objets ». Dans *COmpression et REprésentation des Signaux Audiovisuels (CORESA)*. Rennes, France. **2005**, accepté pour publication.

2. Zlatoff, N., Tellez, B. et Baskurt, A. « Vision Gestalt et connaissance : une approche générique à l'interprétation d'images ». Dans *COmpression et REprésentation des Signaux Audiovisuels (CORESA)*. pages 2355–2358. Lille, France. **2004**.
3. Zlatoff, N., Tellez, B. et Baskurt, A. « Image Understanding Using Domain Knowledge ». Dans *RIAO, Recherche d'Information Assistée par Ordinateur*. pages 277–290. Avignon, France. **2004**.

B

Théorie de l'évidence de Dempster-Shafer

Basée sur les travaux de Dempster, formalisée par Shafer (1976), la théorie de l'évidence est une extension du cadre probabiliste. Elle autorise en particulier une modélisation particulièrement fine de l'incertain et s'attache à décrire la notion de *croyance* en un évènement. En outre, elle est particulièrement adaptée à la problématique de fusion de données : étant donnés plusieurs points de vue sur une hypothèse, comment en déduire un point de vue unique, qui prend en compte tous les autres ?

B.1 Jeu de masses et fonction de croyance

B.1.1 Notations et définitions

Soit Θ un ensemble de n hypothèses mutuellement exclusives $\{H_1, H_2, \ldots, H_n\}$. Θ est appelé cadre de discernement. On suppose que parmi Θ une et une seule hypothèse est vérifiée. C'est l'hyothèse du monde fermé.

Etant donné une partie A de Θ, on note $\overline{A}$ son complément dans Θ. Ainsi, l'évènement $\overline{A}$ désigne l'évènement contraire de A.

On note 2^Θ l'ensemble des parties de Θ.

B.1.2 Jeu de masses

Le point de départ de Shafer dans la théorie de l'évidence a été de modéliser l'incertain. Pour ceci, il revient aux bases de la théorie des probabilités classiques :

Il considère deux hypothèses H_1 et $\overline{H_1}$ de Θ et une fonction de probabilité p. Cette dernière ne peut répartir sa probabilité que de manière exclusive sur H_1 ou sur $\overline{H_1}$. Ceci signifie que si $p(H_1) = x$ alors $p(\overline{H_1}) = 1 - x$.

Shafer considère qu'il serait pertinent, pour modéliser une croyance, que celle-ci puisse allouer une probabilité basique non seulement aux hypothèses prises séparemment (H_1, $\overline{H_1}$), mais plus généralement à toute partie de Θ (par exemple $H_1 \cup \overline{H_1}$) si je ne dispose de connaissances suffisantes que sur cette partie.

Ainsi, on définit un jeu de masses comme une fonction $m : 2^\Theta \rightarrow [0\ 1]$, qui vérifie :

$$m(\oslash) = 0 \tag{B.1}$$

$$\sum_{A \subset \Theta} m(A) = 1 \tag{B.2}$$

$m(A)$ représente la croyance exacte, basique que quelqu'un alloue à l'évènement A. Modéliser un jeu de masses consiste donc à répartir toute la croyance disponible (de valeur posée égale à 1 suivant la convention de l'équation B.2) sur l'ensemble des sous-parties de Θ, en fonction de nos observations et connaissances a priori.

Une partie qui se voit affecter une masse non nulle est appelée élément focal. Dans l'hypothèse du monde fermé, il est impossible d'allouer une masse non nulle à l'ensemble vide $\oslash$. Cela signifierait que l'on alloue une croyance au fait qu'aucune hypothèse n'est vraie, ce qui contredit l'existence même du monde fermé.

B.1.3 Fonction de croyance

Nous venons de voir que la masse $m(A)$ représente la croyance *exacte* que quelqu'un porte en A. Ceci est différent, dans la théorie de l'évidence, de la croyance *totale* que l'on porte sur A. En effet, considérons le cadre de discernement suivant : $\Theta = \{H_1, H_2, H_3\}$. Supposons que nos connaissances a priori sur Θ nous conduisent à poser :

$$m(H_1) = x_1 \tag{B.3}$$

$$m(H_1 \cup H_2) = y \tag{B.4}$$

$$m(\Theta) = 1 - x_1 - y \tag{B.5}$$

L'évènement $H_1 \cup H_2$ possède une masse (croyance exacte) de y. Néanmoins, puisque l'évènement H_1 implique $H_1 \cup H_2$, sa croyance exacte s'ajoute à la croyance totale en $H_1 \cup H_2$. Cet effet de transfert de croyance via les masses est modélisé par

la notion de fonction de croyance. On définit une fonction $Bel : 2^{\Theta} \rightarrow [0\ 1]$ à partir du jeu de masses m avec :

$$\forall A \subset \Theta,\ Bel(A) = \sum_{B \subset A} m(B) \tag{B.6}$$

$Bel(A)$ représente ainsi la croyance *totale* que l'on porte en A. Elle est obtenue en somment les croyances exactes m sur toutes les sous-parties de A.

On peut montrer qu'une fonction Bel est une fonction de croyance si et seulement si :

$$Bel(\oslash) = 0 \tag{B.7}$$
$$Bel(\Theta) = 1 \tag{B.8}$$

pour tout entier k et toute collection $A_1, A_2, , A_k$ de parties de Θ :

$$Bel(\bigcup_{i=0}^{k} A_i) \leq \sum_{I \subset \{1,\ldots,k\}\ I \neq \oslash} (-1)^{card(I)+1} Bel(\bigcap_{i \in I} A_i) \tag{B.9}$$

La première condition stipule que l'on ne peut allouer de croyance à l'évènement vide. On retrouve la notion de monde fermé. La seconde condition normalise tout la croyance disponible à 1. Enfin, et plus remarquable, la dernière condition est une extension de la formule du crible de Poincarré pour les probabilités. Dans sa version à deux parties A_1, A_2, elle s'écrit :

$$Bel(A_1 \cup A_2) \leq Bel(A_1) + Bel(A_2) - Bel(A_1 \cap A_2) \tag{B.10}$$

C'est donc une bien une extension de la formule du crible qui, pour une probabilité p s'écrit :

$$p(A_1 \cup A_2) = p(A_1) + p(A_2) - p(A_1 \cap A_2) \tag{B.11}$$

Formellement, la théorie de l'évidence se construit donc en modifiant le troisième axiome des probabilités.

B.2 Exemples de modélisation

En pratique, pour définir une fonction de croyance Bel, on passe par son jeu de masses associé m. Nous présentons maintenant divers exemples de jeux de masses.

B.2.1 Ignorance totale

$$m(\Theta) = 1 \tag{B.12}$$
$$\forall A \subset \Theta, A \neq \Theta \; m(A) = 0 \tag{B.13}$$

Toute la croyance disponible est concentrée en Θ. Les connaissances dont on dispose ne permettent pas de placer une portion de croyance sur une sous-partie de Θ et donc de raffiner la description.
D'après la condition de normalisation B.2, on peut représenter graphiquement les jeux de masses m sur un axe unité. Le cas de l'ignorance est illustrée figure B.1

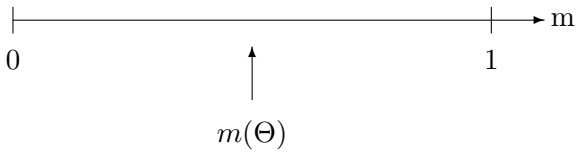


FIG. B.1 – Jeu de masses de l'ignorance totale.

On a alors :

$$Bel(\Theta) = \sum_{B \subset \Theta} m(B) = m(\Theta) = 1 \tag{B.14}$$

B.2.2 Certitude

$$\exists H_i \in \Theta, \; m(H_i) = 1 \tag{B.15}$$
$$\forall A \subset \Theta, A \neq H_i \; m(A) = 0 \tag{B.16}$$

Cette fois, la croyance est toute entière répartie sur une hypothèse de Θ et exprime donc la certitude quant à sa réalisation.

Par suite :

$$Bel(H_i) = \sum_{B \subset H_i} m(B) = m(H_i) = 1 \tag{B.17}$$

B.2.3 Fonction de croyance à support simple

Ces fonctions de croyance ont très souvent été utilisées dans nos modélisations. Nous les détaillons donc maintenant.

$$\exists A \subset \Theta, A \neq \Theta \; m(A) = x \tag{B.18}$$
$$m(\Theta) = 1 - x \tag{B.19}$$

Une portion x de croyance est affectée à l'hypothèse A, et tout le reste est porté sur l'incertitude : bien que le point de vue modélisé croit en A au degré x, il ne croit pas pour autant en l'hypothèse contraire $\overline{A}$ au degré $1 - x$. Cette croyance restante est affectée à l'incertitude Θ. En outre, $m(\overline{A}) = 0$. La situation est illustrée graphiquement à la figure B.2.

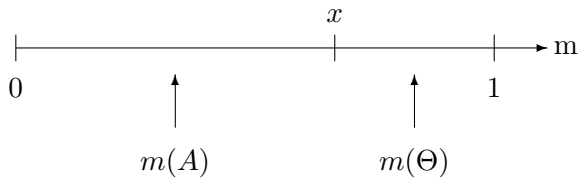


FIG. B.2 – Jeu de masses correspondant à une fonction de croyance à support simple.

On a alors :

$$Bel(A) = \sum_{B \subset A} m(B) = m(A) = x \tag{B.20}$$

Ceci signifie que la croyance totale et exacte en A sont toutes deux égales à x. Par souci de simplicité, nous n'avons donc pas insisté dans le manuscrit sur la différence entre croyance exacte et totale.

$$Bel(\Theta) = \sum_{B \subset \Theta} m(B) = m(A) + m(\Theta) = 1 \tag{B.21}$$

On retrouve donc la condition de normalisation. De plus :

$$\forall C \subset A, \; Bel(C) = \sum_{B \subset C} m(B) = 0 \tag{B.22}$$
$$\forall C \subset \Theta, A \subset C, \; Bel(C) = \sum_{B \subset C} m(B) = m(A) = x \tag{B.23}$$

Ainsi, cette fonction de croyance soutient toute hypothèse C qui impliquée par A (équation B.23), mais aucune autre plus fine que A (équation B.22).

B.2.4 Jeu de masses bayesien

$$\forall H_i \in \Theta,\ m(H_i) = x_i \neq 0 \tag{B.24}$$
$$\forall A \subset \Theta, A \neq H_i\ m(A) = 0 \tag{B.25}$$

Chaque hypothèse de Θ se voit affecter une croyance non nulle et aucune autre partie de Θ ne se voit attribuer une croyance. Le cas à deux hypothèses est représenté à la figure B.3.

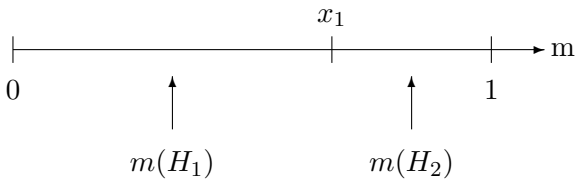


FIG. B.3 – Jeu de masses bayésien sur deux hypothèses $\Theta = \{H_1, H_2\}$.

On a alors :

$$\forall H_i \in \Theta,\ Bel(H_i) = \sum_{B \subset H_i} m(B) = m(H_i) = x_i \tag{B.26}$$
$$\tag{B.27}$$

La fonction Bel est alors équivalente à une fonction de probabilité classique p. Par exemple, pour le cas de la figure B.3 :

$$Bel(\overline{H_1}) = \sum_{B \subset \overline{H_i}} m(B) = m(H_2) = 1 - m(H_1) = 1 - Bel(\overline{H_1}) \tag{B.28}$$

B.3 Règle de combinaison

Une des principaux intérêts de la théorie de l'évidence est, étant donnés plusieurs point de vue différents sur un cadre de discernement Θ, de pouvoir en dériver un point de vue unique, qui prend en compte tous les autres. Cette problématique se

rencontre fréquemment en vision ou en indexation, lorsque l'on dispose de plusieurs descripteurs pour conclure quant à une hypothèse. En pratique, chaque point de vue est modélisé par une fonction de croyance et la combinaison consiste à en calculer une somme orthogonale, que nous allons maintenant expliciter.

B.3.1 Principe

Soit m_1 et m_2 deux jeux de masses respectivement associés aux fonctions de croyance Bel_1 et Bel_2, sur le même cadre de discernement Θ. On note A_i les éléments focaux de Bel_1 et B_j ceux de Bel_2.

Si le nombre $k = \sum_{A_i \cap B_j = \oslash} m_1(A_i)m_2(B_j) < 1$, alors la fonction $m : 2^\Theta \rightarrow [0\ 1]$ définie par :

$$m(\oslash) = 0 \tag{B.29}$$

$$\forall C \subset \Theta, C \neq \oslash,\ m(C) = \frac{\sum\limits_{A_i \cap B_j = C} m_1(A_i)m_2(B_j)}{1-k} \tag{B.30}$$

est un jeu de masses. Sa fonction de croyance associée Bel est appelée somme orthogonale de Bel_1 et Bel_2. Elle représente la croyance combinée sur le cadre de discernement Θ, étant donnés les deux points de vue modélisés par Bel_1 et Bel_2.

Le nombre k est appelée mesure de conflit et tend vers 1 lorsque les fonctions de croyance sont incompatibles, c'est-à-dire soutiennent des hypothèses contraires. Dans ce cas, les résultats obtenus par la règle de combinaison (équation B.30) ne sont pas toujours représentatifs. En particulier, lorsque $k = 1$, le conflit est total et la somme orthogonale n'existe pas.

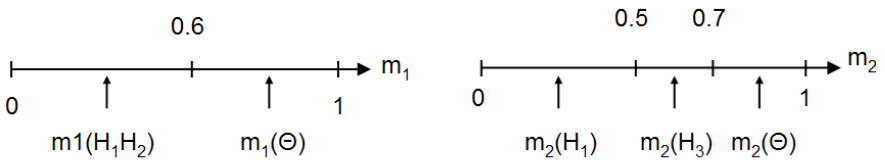
B.3.2 Exemple d'application

B.3.2.1 Cadre de discernement et jeux de masses

Soit $\{H_1, H_2, H_3\}$. Considérons les deux jeux de masses m_1 etm_2 de la figure B.4. m_1 alloue une croyance sur l'hypothèse composée $H_1 \cup H_2$ uniquement $(0, 6)$ et donc place tout le reste $(0, 4)$ sur l'incertitude Θ. m_2 alloue une croyance sur H_1 d'une part $(0, 5)$, H_3 d'autre part $(0, 2)$. Le reste $(0, 3)$ allant encore une fois à Θ. Ainsi, aucune croyance n'est allouée par m_2 à H_2.

B.3.2.2 Fonctions de croyance associées

On peut alors calculer la fonction de croyance Bel_1 associée à m_1 :



(a) (b)

FIG. B.4 – Exemples de jeux de masses.

$$Bel_1(H_1) = \sum_{B \in H_1} m_1(B) = 0 \tag{B.31}$$

$$Bel_1(H_2) = \sum_{B \in H_2} m_1(B) = 0 \tag{B.32}$$

$$Bel_1(H_3) = \sum_{B \in H_3} m_1(B) = 0 \tag{B.33}$$

$$Bel_1(H_1H_2) = \sum_{B \in H_1H_2} m_1(B) = m_1(H_1H_2) = 0.6 \tag{B.34}$$

$$Bel_1(\Theta) = \sum_{B \in \Theta} m_1(B) = 1 \tag{B.35}$$

$$\tag{B.36}$$

De même pour la fonction de croyance Bel_2 associée à m_2 :

$$Bel_2(H_1) = \sum_{B \in H_1} m_2(B) = m_2(H_1) = 0.5 \tag{B.37}$$

$$Bel_2(H_2) = \sum_{B \in H_2} m_2(B) = 0 \tag{B.38}$$

$$Bel_2(H_3) = \sum_{B \in H_3} m_2(B) = m_2(H_3) = 0.2 \tag{B.39}$$

$$Bel_2(H_1H_2) = \sum_{B \in H_1H_2} m_2(B) = m_2(H_1) + m_2(H_2) = 0.5 \tag{B.40}$$

$$Bel_2(\Theta) = \sum_{B \in \Theta} m_2(B) = 1 \tag{B.41}$$

$$\tag{B.42}$$

B.3.2.3 Combinaison de Dempster

La règle de combinaison peut être illustrée graphiquement par la figure B.5 :

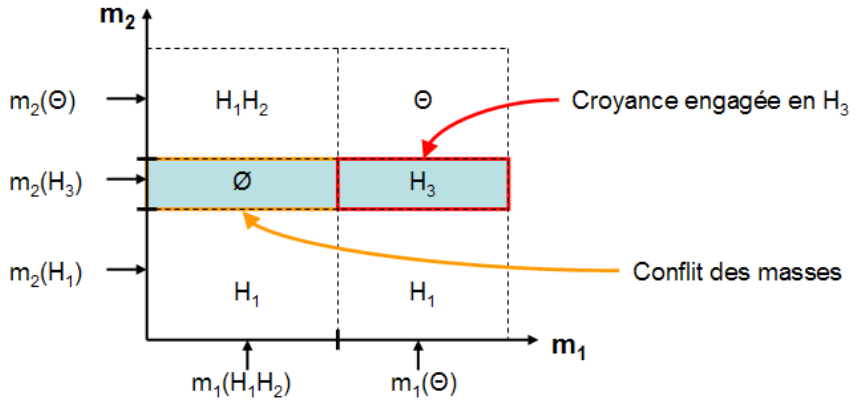


FIG. B.5 – Combinaison des deux jeux de masses de la figure B.4. La croyance engagée en chaque cas est représentée par l'aire de la région associée.

Elle donne lieu à six cas possibles. La croyance résultante affectée à chacun des cas est représentée graphiquement par l'aire correspondante. Toutefois, il faut commencer par calculer la valeur de conflit k :

$$k = \sum_{A_i \cap B_j = \emptyset} m_1(A_i) m_2(B_j) \tag{B.43}$$

$$= m_1(H_1H_2).m_2(H_3) \tag{B.44}$$

$$= 0.6 \cdot 0.2 = 0.12 \tag{B.45}$$

On peut alors en déduire les masses résultantes m allouées à chacun des 5 autres cas :

$$m(H_3) = \frac{m_2(H_2)m_1(\Theta)}{1-k} = \frac{0.2 \cdot 0.4}{1-0.12} = 0.09 \tag{B.46}$$

$$m(H_1) = \frac{0.6 \cdot 0.5 + 0.4 \cdot 0.5}{1-0.12} = 0.57 \tag{B.47}$$

$$m(H_1H_2) = \frac{0.6 \cdot 0.3}{1-0.12} = 0.20 \tag{B.48}$$

$$m(\Theta) = \frac{0.4 \cdot 0.3}{1-0.12} = 0.14 \tag{B.49}$$

Puis, la fonction de croyance associée *Bel* :

$$Bel(H_1) = \sum_{B \in H_1} m(B) = m(H_1) = 0.57 \tag{B.50}$$

$$Bel(H_2) = \sum_{B \in H_2} m(B) = 0 \tag{B.51}$$

$$Bel(H_3) = \sum_{B \in H_3} m(B) = m(H_3) = 0.09 \tag{B.52}$$

$$Bel(H_1 H_2) = \sum_{B \in H_1 H_2} m(B) = m(H_1 H_2) = 0.20 \tag{B.53}$$

$$Bel(\Theta) = \sum_{B \in \Theta} m(B) = 1 \tag{B.54}$$

$$\tag{B.55}$$

B.3.2.4 Discussion

On peut alors comparer qualitativement la fonction de croyance obtenue Bel avec les deux fonctions de croyance de départ Bel_1 et Bel_2.

On peut déjà noter qu'aucune des deux fonctions de croyance Bel_1 et Bel_2 ne soutenait au départ l'hypothèse H_2. Il est donc assez intuitif de retrouver pour Bel une croyance associée en H_2 nulle.

Plus intéressant, on constate que $Bel(H_1)$ s'est renforcée pendant le processus. Elle est en effet supérieure à l'issue de la combinaison à la plus forte des valeurs allouées par Bel_1 ou Bel_2. Ceci peut être expliqué aini : Bel_1 ne soutenait aucunement H_1. Toutefois, elle soutenait l'union $H_1 H_2$. De l'autre côté, Bel_2 agissait de manière similaire. Ainsi, une partie de cette croyance a été transférée pour Bel sur H_1, moyennant un doute résiduel Θ.

Il faut voir que concernant l'hypothèse H_3, le même processus n'a pas pu agir. En effet, aucune des deux fonctions Bel_1 et Bel_2 ne soutenait une hypothèse composée qui aurait inclus H_3. Il en résulte, à l'issue du processus, une croyance combinée inférieure à la plus forte des deux au départ.