



**HAL**  
open science

# Configurations spatiales et segmentation pour la compréhension de scènes, application à la ré-identification

Robin Deléarde

► **To cite this version:**

Robin Deléarde. Configurations spatiales et segmentation pour la compréhension de scènes, application à la ré-identification. Vision par ordinateur et reconnaissance de formes [cs.CV]. Université Paris Cité, 2022. Français. NNT : 2022UNIP7020 . tel-04089313

**HAL Id: tel-04089313**

**<https://theses.hal.science/tel-04089313>**

Submitted on 4 May 2023

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

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

Université Paris Cité, Laboratoire d'Informatique Paris Descartes (LIPADE, URP 2517)  
École Doctorale Informatique, Télécommunications et Électronique (EDITE, ED 130)  
Magellium, avec le soutien de l'Agence de l'Innovation de Défense (AID)

---

# Configurations spatiales et segmentation pour la compréhension de scènes, application à la ré-identification

---

Par ROBIN DELÉARDE

Thèse de doctorat en :

INFORMATIQUE

Spécialité :

TRAITEMENT DU SIGNAL ET DES IMAGES

Dirigée par :

LAURENT WENDLING

Présentée et soutenue publiquement le 12/12/2022

## Membres du jury

MICHEL CRUCIANU	Professeur, CEDRIC, CNAM	Rapporteur
VALÉRIE GOUET-BRUNET	Directrice de recherche, LaSTIG, IGN	Rapporteuse
CATHERINE ACHARD	Professeure, ISIR, Sorbonne Université	Examinatrice
STÉPHANE HERBIN	Directeur de recherche, DTIS, ONERA	Examinateur
LAURENT WENDLING	Professeur, LIPADE, Université Paris Cité	Directeur de thèse
CAMILLE KURTZ	Maître de conférences, LIPADE, Université Paris Cité	Co-encadrant
THOMAS RISTORCELLI	Ingénieur et responsable d'unité, Magellium	Invité
VINCENT ROPERT	Ingénieur, Direction Générale de l'Armement (DGA)	Invité



# Résumé

La modélisation de la configuration spatiale des objets d'une image est un sujet encore peu abordé à ce jour, y compris dans les approches les plus modernes de vision par ordinateur comme les réseaux convolutionnels (CNN). Pourtant, il s'agit d'un aspect essentiel de la perception des scènes, et l'intégrer dans les modélisations devrait bénéficier à de nombreuses tâches du domaine, en contribuant à combler le "fossé sémantique" entre l'image numérique et l'interprétation de son contenu. Ainsi, cette thèse a pour objet l'amélioration des techniques de modélisation de la configuration spatiale, afin de l'exploiter dans des systèmes de description et de reconnaissance.

Dans un premier temps, nous nous sommes penchés sur le cas de la configuration spatiale entre deux objets, en proposant une amélioration d'un descripteur existant. Ce nouveau descripteur appelé "bandeau de forces" est une extension de l'histogramme du même nom à tout un panel de forces, ce qui permet de mieux décrire les configurations complexes. Nous avons pu montrer son intérêt pour la description de scènes, en apprenant à classifier automatiquement des relations en langage naturel à partir de paires d'objets segmentés. Nous avons alors abordé la problématique du passage à des scènes comportant plusieurs objets, proposant une approche par objet en confrontant chaque objet à l'ensemble des autres, plutôt qu'en ayant un descripteur par paire.

Dans un second temps, le contexte CIFRE nous a amenés à traiter une application au problème de la ré-identification de scènes ou d'objets, tâche qui s'apparente à la reconnaissance fine à partir de peu d'exemples. Pour cela, nous nous basons sur une approche traditionnelle en décrivant les constituants de la scène par différents descripteurs dédiés à des caractéristiques spécifiques, comme la couleur ou la forme, auxquelles nous ajoutons la configuration spatiale. La comparaison de deux scènes se fait alors en appariant leurs constituants grâce à ces caractéristiques, en utilisant par exemple l'algorithme hongrois. Différentes associations de caractéristiques peuvent être considérées pour l'appariement et pour le calcul du score final, selon les invariances présentes et recherchées.

Pour chacun de ces deux sujets, nous avons été confrontés aux problèmes des données et de la segmentation. Nous avons alors généré et annoté un jeu de données synthétiques, et exploité deux jeux de données existants en les segmentant, dans deux cadres différents. La première approche concerne la segmentation objet-fond et se place dans le cas où une détection est disponible, ce qui permet d'aider la segmentation. Elle consiste à utiliser un modèle existant de segmentation globale, puis à exploiter la détection pour sélectionner le bon segment, à l'aide de plusieurs critères géométriques et sémantiques. La seconde approche concerne la décomposition d'une scène ou d'un objet en parties et se place dans le cas non supervisé. Elle se base alors sur la couleur des pixels, en utilisant une méthode par *clustering* dans un espace de couleur adapté, comme le cône HSV que nous avons utilisé.

Tous ces travaux ont permis de montrer la possibilité d'utiliser la configuration spatiale pour la description de scènes réelles contenant plusieurs objets, ainsi que dans une chaîne de traitements complexe comme celle utilisée pour la ré-identification. En particulier, l'histogramme de forces a pu être utilisé pour cela, ce qui permet de profiter de ses bonnes performances, en utilisant une méthode de segmentation adaptée au cas d'usage pour traiter des images naturelles.



---

**Mots-clefs :**

configuration spatiale ; compréhension de scènes ; relations spatiales ; histogramme de forces ; graphes relationnels ; segmentation d'images ; espaces de couleurs ; ré-identification ; reconnaissance fine

# Abstract

## **Spatial configurations and segmentation for scene understanding, application to re-identification**

Modeling the spatial configuration of objects in an image is a subject that is still little discussed to date, including in the most modern computer vision approaches such as convolutional neural networks (CNN). However, it is an essential aspect of scene perception, and integrating it into the models should benefit many tasks in the field, by helping to bridge the “semantic gap” between the digital image and the interpretation of its content. Thus, this thesis aims to improve spatial configuration modeling techniques, in order to exploit it in description and recognition systems.

First, we looked at the case of the spatial configuration between two objects, by proposing an improvement of an existing descriptor. This new descriptor called “force banner” is an extension of the histogram of the same name to a whole range of forces, which makes it possible to better describe complex configurations. We were able to show its interest in the description of scenes, by learning to automatically classify relations in natural language from pairs of segmented objects. We then tackled the problem of the transition to scenes containing several objects and proposed an approach per object by confronting each object with all the others, rather than having one descriptor per pair.

Secondly, the industrial context of this thesis led us to deal with an application to the problem of re-identification of scenes or objects, a task which is similar to fine recognition from few examples. To do so, we rely on a traditional approach by describing scene components with different descriptors dedicated to specific characteristics, such as color or shape, to which we add the spatial configuration. The comparison of two scenes is then achieved by matching their components thanks to these characteristics, using the Hungarian algorithm for instance. Different combinations of characteristics can be considered for the matching and for the final score, depending on the present and desired invariances.

For each one of these two topics, we had to cope with the problems of data and segmentation. We then generated and annotated a synthetic dataset, and exploited two existing datasets by segmenting them, in two different frameworks. The first approach concerns object-background segmentation and more precisely the case where a detection is available, which may help the segmentation. It consists in using an existing global segmentation model and exploiting the detection to select the right segment, by using several geometric and semantic criteria. The second approach concerns the decomposition of a scene or an object into parts and addresses the unsupervised case. It is based on the color of the pixels, by using a clustering method in an adapted color space, such as the HSV cone that we used.

All these works have shown the possibility of using the spatial configuration for the description of real scenes containing several objects, as well as in a complex processing chain such as the one we used for re-identification. In particular, the force histogram could be used for this, which makes it possible to take advantage of its good performance, by using a segmentation method adapted to the use case when processing natural images.

---

**Keywords :**

artificial vision; image understanding; spatial configuration; spatial relations; force histogram; relational graphs; image segmentation; color spaces; re-identification; fine-grained recognition

---

*À Gérard Le Chaix, un chef qu'on n'oublie pas.  
À mes parents et grands-parents.*

---



# Remerciements

Je tiens à remercier le LIPADE et l'équipe SIP de m'avoir accueilli pendant ces trois années. En particulier, je remercie mes encadrants pour leur présence et leurs encouragements : Laurent Wendling mon directeur de thèse et Camille Kurtz mon co-encadrant.

Je remercie Michel Crucianu et Valérie Gouet-Brunet d'avoir accepté d'être rapporteurs de ce manuscrit et d'y avoir accordé une attention particulière, ainsi que Catherine Achard et Stéphane Herbin d'avoir accepté de faire partie de mon jury et d'avoir porté de l'intérêt à mes travaux également.

Je remercie l'entreprise Magellium pour avoir mis en place et financé cette thèse jusqu'au bout, ainsi que l'AID pour l'avoir soutenue. Merci également à l'EDITE, au CED Université Paris Cité et à Aurélie pour leurs contributions dans l'administration de cette thèse.

Merci évidemment à mes collègues de bureau, avec qui j'ai pu passer de bons moments pas toujours très productifs : en particulier Olivier, Mohamed, Thibault, Basile, Zhuxian et Amine, avec qui j'ai passé le plus de temps ; mais aussi Qinghe, Christian, Guillaume, Nathan, Rebecca, Lucrezia et François. Merci aussi aux autres membres de l'équipe SIP pour leur bienveillance : Nicole, Nicolas, Sylvain et Florence, avec une pensée pour Georges qui était présent pour le lancement de ma thèse.

Je souhaite aussi remercier tous ceux qui m'ont fait aimer le monde de la recherche, en particulier mes encadrants et anciens collègues de l'ONERA et du CNAM : Stéphane, Bertrand, Fabrice, Aurélien, Martial, Maxime, Hicham, Thibaut, Michel, Marin, Nicolas, Andrey, Francesco, Philippe, Raphaël, Clément, Abdelbadie, Pierre-Henri et Wafa, pour ne citer qu'eux.

Merci également à ceux qui m'ont aidé et aiguillé dans mon parcours, notamment Thierry Coupris, Michel Crucianu, Stéphane Herbin et Vincent Ropert.

Et surtout merci à mes proches et amis qui m'ont porté, soutenu et encouragé.



# Table des matières

<b>Résumé</b>	<b>i</b>
<b>Abstract</b>	<b>iii</b>
<b>Remerciements</b>	<b>vii</b>
<b>Table des matières</b>	<b>ix</b>
<b>Liste des figures</b>	<b>xiii</b>
<b>Liste des tableaux</b>	<b>xv</b>
<b>Liste des sigles et acronymes</b>	<b>xvii</b>
<b>Introduction</b>	<b>1</b>
Contexte . . . . .	3
Compréhension de scènes et applications . . . . .	3
Apprentissage profond et apprentissage traditionnel . . . . .	4
Prise en compte de la configuration spatiale . . . . .	6
Positionnement de la thèse . . . . .	6
Modélisation de la configuration spatiale . . . . .	6
Application à la ré-identification . . . . .	6
Organisation du document . . . . .	7
<b>I Modélisation de configurations spatiales</b>	<b>9</b>
<b>1 Introduction à la modélisation de configurations spatiales</b>	<b>11</b>
1.1 Définition du sujet et problématiques . . . . .	12
1.1.1 Contexte . . . . .	12
1.1.2 Configurations et relations spatiales . . . . .	12
1.1.3 Applications et jeux de données . . . . .	20
1.1.4 Axes de recherche de la littérature . . . . .	23
1.1.5 Défis et positionnement des travaux . . . . .	25
1.2 Génération de données . . . . .	27
1.2.1 Données pour les relations spatiales entre deux objets . . . . .	27
1.2.2 Données pour la configuration spatiale d'une scène ou d'un objet . . . . .	29



<b>2</b>	<b>Configuration spatiale entre deux objets</b>	<b>35</b>
2.1	État de l'art . . . . .	36
2.1.1	Descripteurs de relations spatiales . . . . .	36
2.1.2	Évaluation de relations spatiales en langage naturel . . . . .	40
2.2	Description de position relative et reconnaissance de relations spatiales avec le bandeau de forces . . . . .	44
2.2.1	Approche proposée . . . . .	44
2.2.2	Le bandeau de forces . . . . .	44
2.2.3	Reconnaissance de relations spatiales avec le bandeau de forces . . . . .	51
2.2.4	Expérimentations et résultats . . . . .	53
2.3	Conclusion . . . . .	62
<b>3</b>	<b>Configuration spatiale d'un objet ou d'une scène</b>	<b>65</b>
3.1	Axes de recherche de la littérature . . . . .	65
3.1.1	Généralités . . . . .	65
3.1.2	Graphes relationnels attribués (ARG) . . . . .	66
3.1.3	Sacs de relations et triplets (sujet, relation, objet) . . . . .	69
3.1.4	Vers les configurations spatio-temporelles . . . . .	70
3.2	Descripteurs par parties et comparaison par appariement . . . . .	71
3.2.1	Approche proposée . . . . .	71
3.2.2	Expérimentations . . . . .	73
3.3	Vers la reconnaissance de configurations spatio-temporelles . . . . .	75
3.4	Conclusion . . . . .	76
<b>II</b>	<b>Segmentation pour la décomposition de scène</b>	<b>79</b>
<b>4</b>	<b>Segmentation aidée par des annotations faibles lors de l'inférence</b>	<b>81</b>
4.1	Introduction . . . . .	81
4.1.1	Contexte et positionnement des travaux . . . . .	81
4.1.2	Axes de recherche de la littérature . . . . .	83
4.2	Combinaison de critères visuels et sémantiques pour la sélection de segment . . . . .	84
4.2.1	Approche proposée . . . . .	84
4.2.2	Critères de sélection . . . . .	85
4.2.3	Combinaison de critères . . . . .	87
4.2.4	Expérimentations et résultats . . . . .	87
4.3	Conclusion . . . . .	90
<b>5</b>	<b>Segmentation non supervisée basée sur la couleur</b>	<b>91</b>
5.1	Introduction . . . . .	92
5.2	État de l'art . . . . .	92
5.2.1	Espaces de couleurs numériques polaires (HSX et HCX) . . . . .	93
5.2.2	Segmentation basée sur la couleur . . . . .	94
5.2.3	Classification de données directionnelles ou périodiques . . . . .	97
5.3	Segmentation dans les espaces HSX . . . . .	100
5.3.1	Approches proposées . . . . .	100
5.3.2	Expérimentations et résultats dans l'espace HSV . . . . .	104
5.4	Conclusion . . . . .	108

---

<b>III</b>	<b>Application à la ré-identification</b>	<b>111</b>
<b>6</b>	<b>Ré-identification par appariement de parties</b>	<b>113</b>
6.1	Introduction . . . . .	113
6.1.1	Contexte et positionnement des travaux . . . . .	113
6.1.2	Applications et jeux de données . . . . .	115
6.1.3	Axes de recherche de la littérature . . . . .	117
6.2	Décomposition en parties et comparaison par appariement des parties . . . . .	119
6.2.1	Approche proposée . . . . .	119
6.2.2	Expérimentations et résultats . . . . .	122
6.3	Conclusion . . . . .	124
	<b>Conclusion</b>	<b>127</b>
	Bilan global . . . . .	129
	Retour sur les défis de la configuration spatiale . . . . .	130
	Perspectives globales . . . . .	133
	<b>Annexes</b>	<b>135</b>
A	Historique des espaces de couleurs . . . . .	137
B	Liste des publications . . . . .	141
	<b>Bibliographie</b>	<b>143</b>



# Table des figures

1	Illustrations d'altérations de la configuration spatiale pouvant leurrer un modèle de classification . . . . .	5
1.1	Reconnaissance de configurations spatiales impliquant deux objets . . . . .	13
1.2	Reconnaissance de relations spatiales entre deux objets . . . . .	14
1.3	Exemples de configurations spatiales entre deux objets binaires . . . . .	16
1.4	Exemple de description de scène par des relations spatiales . . . . .	17
1.5	Représentation des angles intervenant dans le point de vue d'acquisition d'une scène .	19
1.6	Exemples de jeux de données dédiés au raisonnement spatial . . . . .	23
1.7	Exemples de différents jeux de données contenant des annotations de relations spatiales	31
1.8	Image synthétique utilisée dans [111] pour la reconnaissance de configurations spatiales entre deux objets . . . . .	32
1.9	Extraits d'une séquence simulée avec <i>Blender</i> . . . . .	32
1.10	Extraits de la séquence <i>CamSeq01</i> : images originales et images segmentées . . . . .	32
1.11	Exemples d'images extraites du jeu de données <i>AeroScapes</i> : images originales et images segmentées . . . . .	33
1.12	Extraits de la séquence 000002 du jeu de données <i>AeroScapes</i> : images originales et segmentations . . . . .	33
1.13	Extraits de la séquence 040000 du jeu de données <i>AeroScapes</i> : images originales et segmentations . . . . .	33
2.1	Illustration des intervalles temporels d'Allen . . . . .	37
2.2	Illustration des relations spatiales topologiques RCC8 . . . . .	37
2.3	Illustration des relations spatiales topologiques et directionnelles selon Hernández . . .	37
2.4	Illustrations de la traduction de descripteurs de position relative en relations spatiales en langage naturel . . . . .	41
2.5	Exemples de paysages flous . . . . .	42
2.6	Illustration du calcul de l'histogramme de forces . . . . .	46
2.7	Comparaison des scores de confiance obtenus avec les forces F0 et F2 . . . . .	47
2.8	Exemple de bandeau de forces discret représenté sous forme de cylindre 3D . . . . .	48
2.9	Exemples d'images et bandeaux de forces des différents jeux de données considérés . .	48
2.10	Illustration des effets de transformations affines de la scène sur l'histogramme de forces	50
2.11	Illustration de la chaîne globale . . . . .	52
2.12	Architecture du réseau <i>SqueezeNet</i> . . . . .	55
2.13	Exemples de bandeaux de forces pour les différentes classes et les différents niveaux d'ambiguïté . . . . .	57
2.14	Courbes d'apprentissage ( <i>loss</i> et <i>accuracy</i> ) . . . . .	59
2.15	Distributions utilisées pour les niveaux de forces : <i>sinh</i> et logarithmique . . . . .	61
2.16	Cartes d'activations obtenues avec différents bandeaux de forces . . . . .	62

3.1	Illustrations de représentation de scène sous forme de graphe relationnel . . . . .	67
3.2	Exemple d'appariement d'objets entre deux scènes . . . . .	67
3.3	Illustration du problème d'appariement de graphes bipartis, ou "affectation" . . . . .	68
3.4	Illustration de la décomposition en histogrammes de forces (FHD) et d'un appariement selon la forme . . . . .	69
3.5	Exemple de graphe spatio-temporel . . . . .	71
3.6	Exemple d'appariement selon les relations à l'aide d'histogrammes de forces "un contre tous" . . . . .	72
3.7	Exemples de descriptions d'objets par histogrammes de forces "un contre tous" . . . . .	73
3.8	Évolution de la similarité des images de la séquence 040000 du jeu de données <i>Aeroscapes</i> selon différentes caractéristiques . . . . .	74
3.9	Évolution de la similarité des images de la séquence 040000 du jeu de données <i>Aeroscapes</i> selon différentes caractéristiques . . . . .	74
3.10	Évolution des histogrammes de forces par paire d'objets sur une scène comportant trois objets, en représentation polaire . . . . .	75
3.11	Évolution des histogrammes de forces par paire d'objets sur une scène comportant trois objets, représentés par des "bandeaux temporels" . . . . .	75
3.12	Évolution temporelle de la similarité entre histogrammes de forces au cours d'une séquence, calculée avec l'indice de Ružička . . . . .	76
4.1	Segmentation d'une image issue du jeu de données PASCAL VOC 2012 . . . . .	85
4.2	Illustration des mesures géométriques utilisées pour le calcul des critères . . . . .	86
4.3	Exemples de segmentations obtenues sur <i>SpatialSense</i> . . . . .	90
5.1	Représentation cubique de l'espace de couleurs RGB . . . . .	102
5.2	Exemples de représentations des espaces de couleurs avec les modèles HSL et HSV . . . . .	103
5.3	Illustration du découpage de l'espace HSI en régions chromatiques et achromatiques . . . . .	104
5.4	Visualisation des composantes HSV d'un extrait d'une image du jeu de données <i>Aeroscapes</i> . . . . .	105
5.5	Sorties de segmentation avec les différentes méthodes proposées . . . . .	107
5.6	Impact du poids de la composante H dans la segmentation dans l'espace HSV conique . . . . .	108
5.7	Visualisation 3D de la segmentation dans le cône HSV . . . . .	109
5.8	Visualisation 2D de la segmentation sur les valeurs de teinte . . . . .	109
6.1	Illustration des tâches de ré-identification à court-terme et long-terme . . . . .	114
6.2	Exemple de segmentation des parties du corps (" <i>human parsing</i> ") . . . . .	116
6.3	Protocole d'acquisition du jeu de données P-DESTRE . . . . .	116
6.4	Synthèse des jeux d'images aériennes pour la ré-identification . . . . .	116
6.5	Illustration de la tâche de reconnaissance à grain fin . . . . .	117
6.6	Exemple de différents détails d'un objet et illustration du processus de vision par saccade . . . . .	118
6.7	Illustration des principaux défis de la ré-identification de personnes . . . . .	119
6.8	Exemples de descriptions d'objets par les caractéristiques des parties . . . . .	122
6.9	Images de test extraites de deux séquences du jeu de données <i>Aeroscapes</i> . . . . .	124
A.1	Évolution des modèles de couleurs principales . . . . .	138
A.2	Illustrations du système de couleurs de Munsell . . . . .	139
A.3	Illustrations du "système naturel de couleurs" (NCS) . . . . .	139
A.4	Illustrations du système de couleurs d'Ostwald . . . . .	139
A.5	Représentation 3D du système de couleurs de Munsell à partir d'ensembles de couleurs de même teinte . . . . .	140

# Liste des tableaux

1.1	Composition des différents jeux de données en termes de complexité . . . . .	28
1.2	Composition des annotations du jeu de données <i>SpatialSense</i> par relation. . . . .	29
2.1	Résultats de classification pour les différentes méthodes . . . . .	58
2.2	Résultats de classification avec le modèle <i>SqueezeNet</i> . . . . .	60
2.3	Résultats de classification pour différentes distributions de forces . . . . .	61
4.1	Scores de segmentation sur l'ensemble de validation de PASCAL VOC . . . . .	89
4.2	Scores de segmentation issus de la littérature sur PASCAL VOC . . . . .	89
6.1	Estimation de la fréquence et de l'importance des variations des caractéristiques pour le suivi et la ré-identification . . . . .	121
6.2	Distance entre le prototype1 et plusieurs images de test avec différents critères . . . . .	124
6.3	Distance entre le prototype2 et plusieurs images de test avec différents critères . . . . .	124



# Liste des sigles et acronymes

- ACP** Analyse en Composantes Principales. 56
- AID** Agence de l'Innovation de Défense. 3
- ARG** *Attributed Relational Graph*. 66, 68, 76, 130
- CAM** *Class Activation Map*. 61, 84
- CBIR** *Content-Based Image Retrieval*. 4, 21, 70, 114, 133
- CIFRE** Convention Industrielle de Formation par la Recherche. i, 129
- CNN** *Convolutional Neural Network*. i, 3–5, 12, 43–45, 47, 51–53, 55–58, 61, 62, 95, 117, 129, 131
- CRF** *Conditional Random Field*. 84, 95
- FHD** *Force Histogram Decomposition*. 68, 69, 71, 76, 77
- HCL** *Hue, Chroma, Lightness*. 94, 96, 109
- HCV** *Hue, Chroma, Value*. 96
- HCX** HCI, HCL ou HCV. 94, 97, 100
- HOG** *Histogram of Oriented Gradients*. 117
- HOI** *Human-Object Interaction*. 43
- HSI** *Hue, Saturation, Intensity*. 93, 97, 102–104
- HSL** *Hue, Saturation, Lightness*. 91–93, 95, 102, 103
- HSV** *Hue, Saturation, Value*. i, 91–93, 98–110, 120, 122, 124, 130, 132
- HSX** HSI, HSL ou HSV. 92–94, 97, 100, 102, 108–110
- IC** *Information Content*. 87
- IHM** Interface Homme-Machine. 21, 22, 28
- IoU** *Intersection over Union*. 40, 73, 88
- IRM** Imagerie à Résonance Magnétique. 22
- MCG** *Multiscale Combinatorial Grouping*. 84, 88
- MIL** *Multiple Instance Learning*. 84
- MLP** *Multi-Layer Perceptron*. 55, 56, 95
- MSER** *Maximally Stable Extremal Regions*. 117
- NLP** *Natural Language Processing*. 4
- OCT** *Optical Coherent Tomography*. 22



- QAP** *Quadratic Assignment Problem.* 67
- RCC** *Region Connection Calculus.* 16, 24, 36
- RGB** *Red, Green, Blue.* 91–94, 96, 97, 100–102, 105, 107
- SIFT** *Scale-Invariant Feature Transform.* 70, 117
- SIG** *Système d’Information Géographique.* 28, 48, 54, 57–60
- SURF** *Speeded-Up Robust Features.* 117
- SVM** *Séparateur à Vaste Marge, ou Support Vector Machine.* 3, 56, 95
- TAL** *Traitement Automatique du Langage.* 4
- TEP** *Tomographie par Émission de Positons.* 22
- TSR** *Triangle Spatial Relationship.* 66
- UCS** *Uniform Color Scale.* 102
- VQA** *Visual Question Answering.* 4, 21, 22, 40
- WSA** *Word Spatial Arrangement.* 70

# Introduction



# Introduction

## Contexte et problématiques

Cette thèse est née d'une volonté de collaboration entre l'entreprise Magellium et le LIPADE sur la modélisation des relations spatiales, notamment pour des applications de reconnaissance, de suivi ou de détection de situations à risque. Elle a été financée par Magellium avec le soutien de l'Agence de l'Innovation de Défense (AID).

## Compréhension de scènes, apprentissage automatique et applications

La compréhension ou interprétation de scènes consiste à fournir une description sémantique du contenu de la scène, à partir d'une image ou d'une vidéo. Elle vise à combler le "fossé sémantique" entre le contenu visuel, c'est-à-dire les pixels constituant les images, et le contenu réel ou sémantique, c'est-à-dire les éléments que représentent ces pixels et leurs relations. C'est un domaine important de la vision par ordinateur qui regroupe de nombreuses tâches, allant de la détection d'objets au raisonnement visuel, en passant par la description en langage naturel. Il repose principalement sur l'apprentissage automatique depuis plusieurs décennies, avec des méthodes telles que les SVM ou les réseaux de neurones, et des applications telles que la détection de visages. Il a vraiment explosé dans les années 2010 avec l'extension des réseaux de neurones convolutifs (CNN) à des images naturelles, après leur introduction deux décennies auparavant [16, 15] et plusieurs améliorations. Cette résurgence a été possible grâce au développement de grandes bases de données annotées (comme *ImageNet* typiquement [10]) et d'architectures plus performantes [14], ainsi qu'à de meilleures performances de calcul (sur cartes graphiques notamment). Leur succès fut tel que désormais la majorité des nouveaux algorithmes d'interprétation d'images reposent sur cette technique.

De la même façon, la description sémantique du contenu s'est développée progressivement dans les tâches de vision par ordinateur. Les premiers traitements ne cherchent pas à la décrire explicitement, mais plutôt à extraire des régions avec des cohérences visuelles, des frontières ou des points d'intérêt. On peut citer notamment les tâches de proposition de régions (ou "blobs" ou plus récemment "superpixels"), de segmentation non-supervisée (objet-fond ou de toute l'image), de détection de contours ou de points d'intérêt. L'étape suivante a alors consisté à décrire ces régions par diverses caractéristiques (de forme, de couleur, de texture...) et représentations, afin de pouvoir les comparer et d'apprendre à reconnaître les objets similaires, donnant naissance à ce qu'on a appelé la reconnaissance de formes ("*pattern recognition*"). Pour réaliser un tel apprentissage, il est nécessaire de disposer d'exemples de chaque classe d'objets d'intérêt, et éventuellement de les ranger par classe manuellement, ce qu'on appelle "annotation". On distingue alors la classification supervisée, où les classes d'apprentissage sont données en entrée, de la classification non supervisée, où le système doit lui-même trouver comment séparer efficacement les données. Dans ce cadre, plusieurs tâches sont possibles sur une nouvelle image, pouvant chacune répondre à une question différente sur le contenu de la scène :

- reconnaissance : quelle est la classe de cet objet? Un score de confiance peut alors être donné pour chaque classe connue;

- détection (et localisation) : où y a-t-il des exemples de cette classe dans l'image ? Ceux-ci sont alors localisés par leurs rectangles englobants ;
- segmentation : quels sont les objets de la scène et leurs limites ? L'image est alors décrite par une classification au niveau pixellique ;
- recherche d'images (CBIR, ou plus simplement "*image retrieval*") : quels sont les exemples les plus proches dans la base ? Ceux-ci peuvent alors être donnés avec un score de similarité.

L'ajout d'une annotation sémantique ("étiquette" ou "*label*") aux différentes classes apprises a aussi permis d'obtenir les premières descriptions automatiques en langage naturel, ainsi que d'établir des correspondances entre image et texte.

En plus de décrire les objets, la recherche en compréhension de scènes s'est ensuite attachée à décrire des configurations particulières de ces objets, telles que des actions ou des émotions, mais aussi les relations entre ces objets, notamment les relations spatiales et les interactions qu'ils peuvent avoir (voir par exemple la Figure 1.7b). La tâche de "*captioning*" consiste à résumer le contenu de l'image en une phrase, en se concentrant sur les éléments essentiels, tandis que la tâche de "*scene parsing*" consiste à décrire de façon exhaustive le contenu de l'image (objets bien délimités, caractéristiques, configurations, interactions). L'utilisation du Traitement Automatique du Langage (TAL), ou *Natural Language Processing* (NLP), permet alors d'aller plus loin dans la compréhension et l'interaction avec le système, en lui permettant d'interpréter des requêtes variées d'un utilisateur, par exemple pour une tâche de *Visual Question Answering* (VQA) et des applications liées à la robotique.

Toutes ces tâches ont de nombreuses applications dans des domaines très variés, à la fois sur des images et des vidéos. On peut citer par exemple, parmi bien d'autres :

- la reconnaissance de caractères pour le traitement automatique des documents ;
- la reconnaissance de visage, d'iris ou d'empreintes pour le contrôle d'identité ;
- la détection de personnes ou de véhicule pour la surveillance aérienne et le renseignement ;
- la classification du sol par imagerie aérienne pour la cartographie, l'agriculture ou la climatologie ;
- la détection et reconnaissance d'animaux dans des images aériennes en bio-statistique ;
- la détection de nuages ou autres phénomènes pour la météorologie ;
- la reconnaissance d'espèce animale ou végétale pour la biologie ;
- la détection d'anomalies et la reconnaissance de pathologies pour la médecine ;
- la reconnaissance d'objets ou de défauts sur des pièces usinées pour l'industrie ;
- la détection de situation anormale pour la conduite autonome ou l'assistance à domicile ;
- la recherche d'objets similaires pour la vente en ligne, la publicité, les réseaux sociaux... ;
- la reconnaissance d'actions pour la vidéo-surveillance ;
- la reconnaissance d'expression du visage pour l'analyse de comportement ;
- etc.

Compte-tenu du contexte industriel et Défense de cette thèse, nous sommes particulièrement intéressés par les applications de détection, reconnaissance et cartographie à partir d'imagerie aérienne ou satellitaire (*remote sensing*). Celles-ci génèrent des contraintes particulières et de nouveaux défis qui suscitent un intérêt croissant, notamment avec les approches par apprentissage profond, comme en témoigne la littérature [11, 20, 2].

## **Apprentissage profond et apprentissage traditionnel**

L'apprentissage profond a permis des avancées considérables dans le domaine de la compréhension de scènes, notamment avec les CNN et plus récemment les transformeurs. Cependant, cette approche a plusieurs défauts majeurs. Tout d'abord, les modèles utilisés demandent de grandes quantités de données annotées pour s'entraîner et s'adapter à une tâche spécifique, ce qui est très coûteux à obtenir, voire impossible pour certaines applications. On constate ainsi depuis quelques années un intérêt croissant pour les méthodes d'apprentissage faiblement supervisé, non supervisé ou auto-supervisé, où l'on dispose de données d'apprentissage mais de peu voire pas d'annotations, et pour les méthodes à partir

de peu de données d'apprentissage (*few-shot/one-shot learning*) voire pas de données du tout (*zero-shot learning*), où l'on ne peut compter que sur des descriptions auxiliaires des classes (en langage naturel par exemple). Toutes ces méthodes ont un intérêt certain dans notre contexte d'application, où il n'est pas facile de disposer de données et encore moins annotées.

Un deuxième défaut important de l'apprentissage profond est la difficulté d'expliquer les décisions qui sont produites par le modèle, du fait du nombre important d'étapes dans le calcul et de paramètres définis automatiquement pour cela. Ainsi, ils sont souvent considérés comme des "boîtes noires", ou plutôt des "boîtes grises" étant donné que les paramètres sont quand même connus, bien qu'ils ne soient pas facilement interprétables. Plusieurs méthodes d'analyse a posteriori (*post-hoc*) ont alors été développées pour mieux comprendre le fonctionnement, les critères de décision et le rôle des paramètres de ces modèles, afin de valider leurs capacités et d'augmenter la confiance qu'un utilisateur peut leur accorder. Ces méthodes peuvent passer par une analyse locale, par exemple en observant les effets de perturbations locales sur les paramètres et la décision, ou par une analyse globale, par exemple avec une visualisation des représentations intermédiaires des données par réduction de la dimensionnalité. Les méthodes locales basées sur l'attention visuelle sont particulièrement utilisées, par exemple pour les transformeurs [13, 3, 1]. Par ailleurs, d'autres méthodes ont cherché à rendre les modèles plus explicables en y intégrant ces approches, comme ProtoPNet [4] qui recherche des ressemblances locales pour générer des prototypes de parties d'objet à l'aide d'un CNN.

Enfin, compte-tenu de la solution qu'ils utilisent pour obtenir l'invariance spatiale, les CNN montrent des limites lorsqu'ils traitent de configurations spatiales variables entre des objets ou des parties d'objet, ce qui peut les empêcher de fournir une description suffisamment fiable et précise de ces configurations et entraîner des erreurs importantes (cf. Figure 1 et Section 1.1.1). Ce problème a notamment été mis en avant dans la proposition des réseaux à capsules [12], mais les implémentations actuelles n'ont pas encore permis d'y répondre.



FIGURE 1 – Illustrations d'altérations de la configuration spatiale pouvant leurrer un modèle de classification.

Face à cela, les méthodes traditionnelles d'apprentissage à partir de descripteurs définis "manuellement" retrouvent tout leur intérêt, après avoir été délaissées suite à l'explosion de l'apprentissage profond par CNN. En effet, ces méthodes ayant moins de paramètres à définir, elles ne nécessitent pas autant de données d'apprentissage, et elles ont l'avantage d'être naturellement interprétables, ce qu'on appelle l'explicabilité par construction. Les descripteurs sont construits grâce aux connaissances du domaine pour mesurer des caractéristiques particulières des objets, comme leur forme, leur couleur, leur texture, ce qui permet d'obtenir des descripteurs génériques pouvant être utilisés pour différentes tâches. La classification par apprentissage pour une tâche donnée se fait alors en apprenant à prédire les classes directement à partir de ces descripteurs, de façon plus ou moins supervisée, en se basant sur l'expertise métier pour savoir quelles caractéristiques utiliser pour une tâche donnée. Ainsi, l'idée est d'apporter de la connaissance sur les images et la tâche dans le modèle, afin de simplifier le problème à traiter par la machine. À noter que les réseaux de neurones profonds permettent également d'obtenir de tels descripteurs génériques (alors appelés "*deep features*"), en utilisant des sorties intermédiaires d'un réseau pré-entraîné pour une tâche suffisamment proche ou large (une classification sur *ImageNet* typiquement), mais leur explicabilité est moins immédiate.

## Prise en compte de la configuration spatiale

Afin de décrire une scène composée de plusieurs objets, ou un objet composé de plusieurs parties, il est utile d'exploiter cette décomposition et de décrire séparément chaque objet ou chaque partie. Différents types de caractéristiques sont exploitables pour cela. Il peut s'agir de caractéristiques liées à une partie, comme sa forme, sa couleur, sa texture, ou encore sa position dans la structure. Mais il peut aussi s'agir de caractéristiques associant plusieurs parties, en décrivant leurs relations. En particulier, la prise en compte de la configuration spatiale est souvent négligée par les approches actuelles, alors qu'elle joue un rôle particulièrement important dans la perception du contenu d'une image. Ces aspects sont en effet très informatifs sur le contenu de l'image, des configurations différentes d'objets pouvant donner lieu à des situations très différentes, et des configurations différentes de parties pouvant constituer des objets très différents. Ainsi, intégrer les relations spatiales dans la description des scènes devrait permettre d'importants progrès dans de nombreuses tâches, en contribuant à combler le "fossé sémantique" entre l'image numérique et la description qu'on pourrait en faire en langage naturel. Par ailleurs, l'ajout de la dimension temporelle avec la vidéo conduit à d'autres possibilités encore, en considérant le mouvement relatif des parties, c'est-à-dire leurs relations spatio-temporelles.

## Positionnement de la thèse

### Modélisation de la configuration spatiale

L'objectif principal de cette thèse est de pouvoir décrire une scène en prenant en compte la configuration spatiale des objets qui la composent, en particulier pour des scènes complexes ou dont on dispose de peu de données annotées. L'idée générale est d'introduire de la connaissance sur la configuration spatiale dans la modélisation, en utilisant des approches "traditionnelles" plus explicables et ne nécessitant pas de lourd apprentissage. De nombreuses tâches seraient alors possibles : détecter les changements entre deux scènes, détecter des anomalies (changements brusques par exemple), reconnaître des configurations connues (par apprentissage typiquement), rechercher des scènes similaires dans une base, suivre ou encore prédire l'évolution de la scène. Les applications sont nombreuses également, que ce soit dans le domaine civil ou militaire : surveillance, détection de menaces, reconnaissance de schémas tactiques, navigation basée vision, navigation autonome, assistance vocale, etc.

Dans un premier temps, nous nous sommes penchés sur des descripteurs de relations spatiales entre deux objets, en proposant une amélioration d'un descripteur existant et en montrant son intérêt pour la description automatique de scène. Puis, afin de décrire et comparer des scènes composées de plusieurs objets, ou des objets composés de plusieurs parties, nous avons exploré plusieurs méthodes de modélisation des éléments et de leurs relations, et proposé une méthode originale pour cela. Par ailleurs, ces descripteurs étant calculés sur des formes binaires, afin de décrire précisément les relations spatiales entre celles-ci, nous avons aussi dû aborder le problème de la segmentation d'objets, que nous avons traité dans deux cadres différents.

### Application à la ré-identification

Le contexte industriel de cette thèse nous a également conduits à étudier une application intéressante l'entreprise Magellium. Plusieurs cas d'usage ont été identifiés au début de la thèse :

- reconnaissance et suivi de cibles ;
- recherche d'image dans une base (*image retrieval*) ;
- localisation (terrestre/aérienne) basée vision, par reconnaissance de l'environnement ;
- prédiction de situations à risque (pour la navigation autonome ou la tenue de situation tactique par exemple), par détection de changement ou apprentissage de configurations spatiales ou spatio-temporelles (ex : stratégie d'attaque).

Le cas d'usage retenu par l'entreprise est celui de la ré-identification, et plus particulièrement de la ré-identification de personnes dans des vidéos aériennes, pour lequel une solution avait déjà été développée mais ne donnait pas satisfaction. Il s'agit d'un cas particulier de reconnaissance de cibles où l'on dispose de quelques exemples d'apprentissage seulement, et où les conditions d'acquisition ne sont pas contrôlées, ce qui signifie que le point de vue peut être différent d'une image à l'autre. De plus, la problématique de temps réel peut aussi être importante afin de traiter "à la volée" les images provenant d'une acquisition vidéo.

Après une étude approfondie de la solution actuelle, nous avons proposé une refonte de celle-ci afin de mieux traiter chaque étape et en particulier la modélisation de la configuration spatiale. Nous verrons ainsi en quoi la configuration spatiale peut être utile dans ce cas d'usage et comment la mettre en œuvre en l'associant à une étape de décomposition de la scène. Pour cela, nous avons été confrontés à deux défis : 1. celui du passage de descriptions de relations entre deux objets (ou parties d'objets) à une description complète de l'agencement de la scène (ou de l'objet), ce qui a nécessité d'explorer de nouvelles modélisations ; 2. celui de la décomposition d'objets en parties, pour lequel nous avons étudié les approches de segmentation basée sur la couleur. La combinaison de ces travaux à un apprentissage de prototypes nous a alors permis d'apporter une solution complète au problème de ré-identification.

## Organisation du document

Le présent document s'articule selon trois parties, comprenant respectivement trois, deux et un chapitres.

La première est consacrée à la description de configurations spatiales. Elle débute dans le Chapitre 1 avec une introduction de cette problématique et de ses applications, et présente les différents jeux de données que nous avons générés et utilisés dans nos travaux. Puis, dans le Chapitre 2, elle se focalise sur la description de configurations spatiales entre deux objets. Après un état de l'art de ce sujet, nous y présentons l'approche que nous avons proposée pour les configurations complexes avec le bandeau de forces. Enfin, elle se conclut avec l'étude dans le Chapitre 3 des descriptions de configurations spatiales de scènes composées de plusieurs objets, ou d'objets composés de plusieurs parties, en donnant un aperçu des méthodes actuelles et en proposant une nouvelle approche par appariement de parties.

La deuxième partie traite quant à elle des problématiques de segmentation que nous avons dû aborder afin d'utiliser les descripteurs précédents sur des images naturelles. Nous avons rencontré deux cas différents où une étape de segmentation était nécessaire : pour extraire des objets à partir d'annotations de détection (rectangle englobant et étiquette) et pour décomposer des objets en parties. Le premier cas nous a conduits à nous intéresser dans le Chapitre 4 à la segmentation objet/fond, et plus spécifiquement au cas où l'on dispose d'annotations faibles lors de l'inférence. Nous y proposons une solution basée sur un modèle pré-entraîné et une sélection de la meilleure sortie en fonction de plusieurs critères. Pour le second cas, nous avons opté pour une segmentation basée sur la couleur, ce qui nous a amenés à explorer dans le Chapitre 5 la définition des espaces de couleurs et en particulier les espaces polaires, afin de développer une méthode cohérente de *clustering* dans un tel espace.

Pour terminer, la troisième partie plus exploratoire concerne l'application à la ré-identification dans le Chapitre 6, combinant les travaux précédents sur la décomposition en parties et l'appariement basé sur la configuration spatiale, parmi d'autres caractéristiques. Elle permet d'illustrer l'intérêt des différentes approches développées, constituant une première preuve de concept pour cette application.

Enfin, les bibliographies sont données par thématique : une pour la première partie, une pour chacun des chapitres 4, 5 et 6, et une globale pour cette introduction et la conclusion.





## **Partie I**

# **Modélisation de configurations spatiales**



# Chapitre 1

## Introduction à la modélisation de configurations spatiales

---

1.1	Définition du sujet et problématiques . . . . .	12
1.1.1	Contexte . . . . .	12
1.1.2	Configurations et relations spatiales . . . . .	12
1.1.2.1	Définitions . . . . .	12
1.1.2.2	Champ lexical de la configuration spatiale . . . . .	14
1.1.2.3	Relations et configurations en langage naturel . . . . .	15
1.1.2.4	Prise en compte du point de vue et de la pose des objets . . . . .	19
1.1.3	Applications et jeux de données . . . . .	20
1.1.3.1	Types de configurations d'intérêt . . . . .	20
1.1.3.2	Tâches applicatives et cas d'usage . . . . .	21
1.1.3.3	Jeux de données . . . . .	22
1.1.4	Axes de recherche de la littérature . . . . .	23
1.1.4.1	Généralités . . . . .	24
1.1.4.2	Descripteurs de configuration spatiale entre deux objets . . . . .	24
1.1.4.3	Description de configuration spatiale d'une scène ou d'un objet . . . . .	25
1.1.4.4	Prise en compte du point de vue . . . . .	25
1.1.5	Défis et positionnement des travaux . . . . .	25
1.2	Génération de données . . . . .	27
1.2.1	Données pour les relations spatiales entre deux objets . . . . .	27
1.2.1.1	Génération et annotation d'un jeu de données synthétique . . . . .	27
1.2.1.2	Découpage et annotation d'une image de télédétection . . . . .	28
1.2.1.3	Segmentation d'un jeu de données réelles annotées . . . . .	28
1.2.2	Données pour la configuration spatiale d'une scène ou d'un objet . . . . .	29
1.2.2.1	Génération de scènes simulées avec <i>Blender</i> . . . . .	29
1.2.2.2	Extraction des objets dans une scène . . . . .	29
1.2.2.3	Décomposition d'objets en parties dans des séquences . . . . .	30

---

## 1.1 Définition du sujet et problématiques

### 1.1.1 Contexte

Depuis de nombreuses années, il est acquis qu'il est important d'exploiter les relations spatiales entre les objets pour mieux comprendre les scènes et mieux les exploiter dans des tâches haut niveau, mais aussi que les descripteurs usuels (de forme, couleur ou texture) ne suffisent pas pour décrire correctement des scènes réelles, par nature plus complexes que celles utilisées pour répondre à une tâche précise. Par rapport aux descriptions usuelles, où l'on cherche typiquement à détecter et reconnaître des objets dans la scène, il s'agit ici d'aller plus loin dans la compréhension de ce qu'elle contient et de ce qu'il s'y passe, de façon à combler le "fossé sémantique" entre les pixels constituant les images et le contenu sémantique qu'ils représentent, et à se rapprocher de la perception humaine [107].

Pourtant, ce sujet est encore peu étudié dans la littérature, et les approches les plus récentes ne considèrent la description des relations que comme un résultat final, utilisant le langage naturel pour l'exprimer, plutôt que comme un autre type de caractéristiques pouvant être utilisé comme entrée dans un système de reconnaissance de formes. Ces informations spatiales devraient compléter les caractéristiques traditionnelles basées sur la forme, la couleur, la texture, mais aussi celles issues des réseaux convolutifs profonds ("*deep features*"), qui ne sont pas toujours suffisantes pour décrire correctement des images composées d'objets avec des configurations spatiales complexes. De plus, il est intéressant de noter que la configuration spatiale est moins sensible aux variations de point de vue que d'autres types de descripteurs [7], à condition d'utiliser des caractéristiques robustes à ces variations. Cela conduit au problème fondamental de la modélisation de la configuration spatiale : étant donnée une image composées de plusieurs objets, comment modéliser efficacement sa configuration spatiale ?

En parallèle, la dernière décennie a vu le développement des réseaux de neurones convolutifs profonds (CNN), avec des résultats exceptionnels dans toutes les tâches d'analyse d'images. Les caractéristiques convolutives extraites grâce aux CNN peuvent inclure un certain contenu spatial s'ils sont entraînés à distinguer des objets avec des configurations spécifiques, mais ils souffrent de limitations importantes, comme détaillé dans l'introduction de ce manuscrit. Premièrement, ils nécessitent de grandes quantités de données d'entraînement, qui sont peu disponibles car elles doivent être annotées manuellement par des humains. Deuxièmement, ils ne sont pas toujours performants pour gérer les informations spatiales comme la distance et l'orientation entre les objets, car ils utilisent un point de vue souvent trop local. En particulier, les étapes de *pooling* font perdre la localisation précise des activations, ce qui affecte la précision spatiale globale. De plus, ils ne peuvent pas gérer naturellement des objets de différentes tailles, car ils sont contraints par la taille de leurs filtres, sauf si une architecture multi-échelles est utilisée. Et enfin, ils sont souvent considérés comme des "boîtes noires", où toutes les caractéristiques sont mélangées et implicites, alors que les systèmes opérationnels ont besoin de décisions explicables. C'est pourquoi il semble préférable de modéliser plus explicitement le contenu spatial, notamment pour les configurations spatiales ambiguës.

### 1.1.2 Configurations et relations spatiales

#### 1.1.2.1 Définitions

On considère une scène comme étant un ensemble d'objets plus ou moins mobiles et déformables, plus ou moins en lien les uns avec les autres, et on considère un objet comme étant un ensemble de parties plus ou moins mobiles et déformables. Dans toute la suite du document, nous parlerons tantôt de scène et tantôt d'objet, les deux pouvant se substituer selon l'application.

La configuration spatiale désigne la manière dont les objets de la scène (ou les parties d'un objet) sont positionnés les uns par rapport aux autres, c'est-à-dire son agencement, sa disposition, ce qui inclut la connaissance de la composition de la scène, mais aussi de sa forme globale voire de celle de

chacun de ses constituants (cf. Section 1.1.2.2). Elle est donc plus large que la simple description de la forme globale, mais aussi que la co-occurrence des parties, qui ne prennent pas en compte la manière dont elles sont organisées. Le terme de configuration est également très proche de celui de structure, mais la structure ne prend pas forcément en compte l'aspect spatial précis, s'attachant davantage à la décomposition hiérarchique des parties. Deux approches sont à distinguer pour la description de configurations spatiales : la description globale de la configuration de la scène et la description des relations spatiales entre ses constituants.

Dans le premier cas, on cherche à décrire la scène ou l'objet en tant qu'ensemble d'éléments, organisés ou pas selon le cas. Pour un objet, la configuration spatiale est intimement liée à la forme, puisqu'elle inclut cette notion : là où la forme se contente de décrire l'aspect global des contours de l'objet, la configuration spatiale considère la décomposition de celui-ci en parties et la position relative de chacune des parties par rapport aux autres, leur agencement. En fait, on parle plus facilement de forme pour un objet, dont les parties sont souvent adjacentes, tandis qu'on parle plutôt de configuration spatiale pour une scène, où les objets sont plus souvent espacés. On peut alors s'intéresser à des configurations caractéristiques, que l'on cherche à reconnaître, comparer ou apparier, comme illustré dans la Figure 1.1 (avec seulement deux objets). Il peut s'agir de configurations simples, dont une description simple en langage naturel existe, comme une forme géométrique (carré, cercle...), ou des configurations plus complexes ou inhabituelles, auquel cas il peut être utile de décomposer le problème en sous-ensembles d'objets et de passer par une modélisation informatique. Des exemples de configurations particulières en langage naturel sont donnés dans la Section 1.1.2.3, tandis que les modèles informatiques de configurations sont décrits dans la Section 1.2.2.



FIGURE 1.1 – Reconnaissance de configurations spatiales impliquant deux objets (source : [111]).

Dans le second cas, on cherche à décrire les relations entre objets plutôt que la configuration globale, en considérant des paires d'objets typiquement, de façon exhaustive ou pour certains objets d'intérêt uniquement, comme illustré dans la Figure 1.2. Une relation est un lien entre plusieurs éléments, qu'on peut appeler des acteurs ou des nœuds, dans une modélisation par un graphe relationnel typiquement. Elles peuvent être orientées, lorsque la position des éléments a une importance, ou non orientées dans le cas contraire. Dans le cas d'une relation orientée entre deux éléments, on utilise les notions de référent et d'argument pour désigner l'élément par rapport auquel on définit la relation et l'élément visé par elle. On parlera alors d'inverse sémantique pour désigner la relation symétrique où le rôle du référent et de l'argument sont inversés. Dans le cas d'une relation orientée impliquant davantage d'objets, plusieurs référents ou arguments peuvent intervenir, selon la définition de la relation.

Par définition, les relations spatiales sont des relations qui font appel à la position relative des objets. À nouveau, la description des relations simples peut se faire avec des relations simples en langage naturel, dont la plupart sont des relations orientées entre deux objets, souvent sous forme de prépositions, comme les relations "à gauche de", "devant" ou "autour de". On obtient alors des descriptions sous la forme de triplets (sujet, relation, objet), le terme de "prédicat" qui est parfois utilisé pour la relation étant grammaticalement plus large en réalité. Pour les situations plus complexes ou inhabituelles, il est alors nécessaire d'utiliser des relations composites comme "à gauche et en-dessous de", ou avec des descriptions plus fines et quantitatives, faisant appel aux mathématiques et à l'informatique. Des

exemples de relations particulières en langage naturel sont donnés dans la Section 1.1.2.3, tandis que les modèles informatiques de description de positions relatives sont décrits dans la Section 1.2.1.



What is the spatial organization of the scene?

*"The dog is to the left of the cat"*

FIGURE 1.2 – Reconnaissance de relations spatiales entre deux objets.

On peut donc distinguer deux types de descriptions liées à la configuration spatiale :

- les descriptions de relations spatiales entre objets, impliquant un nombre restreint d'objets (deux en général),
- les descriptions globales de configurations spatiales de scènes, pouvant comporter un nombre d'objets plus important.

### 1.1.2.2 Champ lexical de la configuration spatiale

Le domaine d'étude de la configuration spatiale a un champ lexical important en langue française, où l'on trouve plusieurs mots aux significations très proches. Nous détaillons ici plusieurs groupes lexicaux : les termes liés à la notion de position d'un objet dans l'espace, ceux liés à la forme d'un objet, pour arriver à ceux liés à l'agencement de plusieurs objets ou parties d'objets. Les définitions indiquées entre guillemets sont celles du Larousse <sup>1</sup>.

Lorsque l'on traite d'information spatiale, la première grandeur que l'on utilise est la position des objets, ou des pixels dans une image, donnée soit avec des coordonnées géographiques dans le monde réel, soit pixelliques dans la prise de vue. Dans notre contexte d'étude, la position est la "place précise occupée dans l'espace par quelque chose, quelqu'un". On parle par exemple de "la position d'une ville sur la carte". Positionner consiste alors à "indiquer les coordonnées géographiques, l'emplacement exact (d'un navire, d'un engin, d'une troupe, etc.)". Le terme "position" est synonyme "d'emplacement" ou de "localisation", mais se distingue du "positionnement" qui est "l'action de positionner" ou le "fait d'être positionné". On trouve également la forme pronominale "se positionner" qui consiste soit à "se placer en un lieu, à un rang précis, déterminé", soit à "se situer, se définir par rapport à quelqu'un, quelque chose", introduisant la notion de relation avec d'autres entités, qui peut être autre que spatiale. La configuration spatiale peut alors être définie comme la position relative des entités, leur "agencement" dans l'espace. Enfin, une autre définition du terme "position" en lien avec nos travaux est celle synonyme de "posture" ou "pose", qui indique en fait une configuration spatiale particulière lorsque l'on parle des parties du corps, comme une posture de yoga.

La notion de configuration est aussi intimement liée à celle de forme. En effet, en langue française, la configuration (du latin ecclésiastique *configurare*, donner une forme) désigne la "forme extérieure d'un ensemble", comme "la configuration d'un pays", tandis que la forme désigne "l'organisation des contours d'un objet", mais aussi "sa structure, sa configuration". De façon étendue, elle désigne l'organisation spatiale des parties d'un ensemble, avec pour synonymes les termes "agencement", "disposition" ou encore "arrangement". On parle par exemple de configuration d'une ville, d'une salle, d'une molécule... Le qualificatif "spatiale", même s'il peut paraître redondant, permet d'insister sur cet aspect, et d'éviter la confusion avec d'autres sens. En anatomie, on parle aussi de "conformation" (du latin *conformatio*), qui est "la manière dont sont assemblées les parties du corps, d'un organe, du squelette". La morphologie quant à elle concerne les sciences biologiques et désigne "la forme de la structure externe des êtres

1. <https://www.larousse.fr/>

vivants", notamment du corps humain. Elle a pris un sens différent en mathématiques et en informatique, où elle considère la structure des objets en termes de connectivité et de topologie, et consiste à les extraire ou améliorer leur manipulation, en leur appliquant diverses opérations comme la fermeture, l'ouverture, l'érosion ou la dilatation. Par ailleurs, en langue allemande, la traduction "*Gestalt*" du mot "forme" a pris un sens particulier en psychologie (cf. ci-dessous).

Disposition, arrangement et agencement ont des significations très proches, de même que les verbes qui s'y rapportent. La disposition (du latin *dispositio*) est définie comme "la manière dont les éléments d'un ensemble ont été disposés", l'action de disposer (du latin *disponere*, avec l'influence de poser) consistant à "arranger des objets d'une certaine manière". L'agencement est "l'état de ce qui est agencé", l'action d'agencer consistant à "arranger un ensemble de sorte que ses éléments soient exactement adaptés les uns aux autres et que le tout réponde au mieux à sa destination". Et l'arrangement désigne "la manière dont les choses sont arrangées", arranger signifiant "mettre des objets, quelque chose dans l'ordre ou la disposition estimés convenables, satisfaisants". Dans l'usage courant, on emploie les termes d'agencement et arrangement uniquement pour des objets inanimés, dont la position ne change pas facilement, et celui de disposition pour tout type d'objets. On parle par exemple de l'agencement d'un appartement, c'est-à-dire de la disposition des pièces dans celui-ci, de l'arrangement d'une coiffure, ou encore de la disposition des sentinelles autour d'un camp. D'autres termes sont aussi en lien avec ces notions, comme les verbes "organiser", "aménager", "ranger", "ordonner", "structurer", "composer", entre autres.

Ainsi, l'expression de "configuration spatiale" que nous introduisons regroupe à la fois les notions de forme et de positions relatives des objets, c'est-à-dire d'agencement. Cependant, contrairement aux termes précédents, celle-ci n'est pas limitée à certains objets inanimés et habituels, ce qui convient davantage à nos travaux sur tous types d'objets, avec des scènes composées d'objets divers et mobiles.

### À propos de la *Gestalt*

*Gestalt* est un mot allemand signifiant "structure, forme, configuration", utilisé tel quel dans d'autres langues (notamment le français et l'anglais) pour désigner l'étude de la perception structurelle en psychologie. Selon le Larousse, c'est le "fait, pour une entité perceptive, d'être traitée par le sujet comme un tout plutôt que comme une juxtaposition de parties". Il a donné son nom à une théorie psychologique et philosophique, appelée gestaltisme ou théorie de la forme, selon laquelle les processus de la perception et de la représentation mentale traitent les phénomènes comme des formes globales plutôt que comme l'addition ou la juxtaposition d'éléments simples. L'idée du gestaltisme est que le cerveau va chercher à structurer les éléments que l'œil perçoit, selon plusieurs lois empiriques. Cette théorie est particulièrement mise en application dans les domaines de l'ergonomie et du *web design*. Les liens entre la *Gestalt* et l'analyse d'images ont notamment été étudiés dans [34] et enseignés pendant plusieurs années dans certaines grandes écoles en France et ailleurs.

#### 1.1.2.3 Relations et configurations en langage naturel

Dans le langage naturel, un certain nombre de mots sont dédiés à la description de relations, et quelques-uns à la description de configurations spatiales. Ils renvoient alors à des relations ou configurations caractéristiques, facilement identifiables, que nous détaillons ici en reprenant les définitions introduites dans la Section 1.1.2.1. Les relations et configurations plus complexes, originales ou ambiguës nécessitent alors d'autres solutions, dont nous donnons un aperçu dans la Section 1.1.4. Nous ne présentons que les mots de la langue française, mais des mots avec le même sens existent aussi dans les autres langues en général. Il serait intéressant néanmoins de mener une étude linguistique approfondie sur ce sujet.



## Relations spatiales entre objets

Dans la vie de tous les jours, nous utilisons les relations spatiales à de nombreuses occasions, que ce soit pour nous repérer, décrire une scène, indiquer l'emplacement d'un objet... Plusieurs études se sont penchées sur les mécanismes qui interviennent pour cela, notamment en sciences cognitives [61, 40, 64], en associant également des études linguistiques [113, 52] pour décrire la sémantique des relations spatiales. En informatique, les premiers travaux formalisant ce type d'information sont ceux de Freeman, qui en 1975 proposa une catégorisation en 13 relations spatiales élémentaires en langage naturel [39]. Ces relations de nature qualitative peuvent être classées selon trois types : les relations topologiques, directionnelles ou distancielles. D'autres relations topologiques peuvent alors être ajoutées, en se basant par exemple sur le modèle RCC [95, 23] (cf. Figure 2.2).

Nous proposons de retenir la liste de relations élémentaires suivante :

- directionnelles : "à gauche de", "à droite de", "au-dessus de", "en-dessous de", "devant" (en 3D), "derrière" (en 3D);
- topologiques : "à l'intérieur de" (ou "dans"), "englobe", "à l'extérieur de", "disjoint de", "recouvre partiellement" (ou "intersecte"), "tangent à" (ou "au contact de");
- distancielles : "proche de" (ou "dans le voisinage de"), "loin de".

Par rapport aux relations proposées par Freeman, nous avons ajouté la relation "englobe", qui est l'inverse sémantique de la relation "à l'intérieur de", la relation "disjoint", qui est plus générale que "à l'extérieur de", et la relation "recouvre" (ou "intersecte"), qui est l'inverse sémantique de la "disjoint" et plus générale que "à l'intérieur de". En revanche nous n'avons pas inclus dans cette liste les relations "à côté de" et "entre", qui sont en fait des combinaisons de deux relations élémentaires : la combinaison de "proche de" et "à droite de" ou "à gauche de" pour la relation "à côté de" (pour un même couple d'objets), et la combinaison de deux relations directionnelles opposées ("à gauche de" et "à droite de", "devant et "derrière" ou "au-dessus de" et "en-dessous de") pour la relation "entre", qui a la singularité d'impliquer trois objets différents. Cette dernière peut aussi être définie par "au milieu de", "de part et d'autre de" ou "à l'opposé de ... par rapport à ..." selon les objets considérés comme référents et arguments. De la même façon, la relation "sur" peut être traduite comme la combinaison des relations "au-dessus de" et "au contact de", pour un même couple d'objets, et la relation "autour de" (ou son inverse sémantique "entouré par") traduit la combinaison de plusieurs relations directionnelles, pouvant impliquer plusieurs objets. Il est possible de combiner d'autres relations pour décrire d'autres situations, sans qu'un mot n'existe forcément pour cela (par exemple : "à gauche et au-dessus de", "loin à droite de", etc.).

Par ailleurs, il est à noter que certaines de ces relations notamment directionnelles ou topologiques peuvent dépendre de la nature et de la pose des objets, aboutissant à une interprétation différente (cf. Section 1.1.2.4). D'autres relations peuvent alors intégrer cette notion de pose, comme les relations "face à", "dos à", "aligné avec", "côte à côte", "face à face", "dos à dos", etc. Enfin, d'autres relations prenant en compte la notion de forme ou d'agencement peuvent aussi être ajoutées, comme les relations "le long de", "parallèle à", "perpendiculaire à", qui s'entendent pour des formes allongées, ou les relations d'enlacement et d'entrelacement pour des formes en spirales. La relation "autour de" peut également prendre en compte la forme lorsqu'elle concerne un seul argument, qui serait alors arrondi.

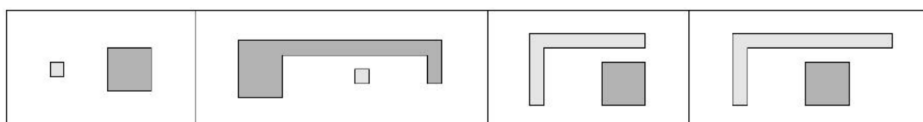


FIGURE 1.3 – Exemples de configurations spatiales entre deux objets binaires (source : [96]). Dans certaines situations, les approches centroïde ou rectangle englobant peuvent entraîner des ambiguïtés.

Ainsi, exceptée la relation "entre" qui implique forcément trois objets, toutes ces relations sont des relations entre deux objets, ou entre deux groupes d'objets. Pour décrire une scène contenant davantage d'objets, il est donc nécessaire de considérer plusieurs paires, ou d'utiliser des descriptions dédiées à la configuration lorsque cela est possible, en précisant alors la position de chacun dans la configuration. Une telle approche qualitative a l'avantage de fournir directement une description en langage naturel, mais elle peut vite devenir complexe. De plus, sa nature "tout ou rien" n'est manifestement pas adaptée pour décrire des configurations complexes ou ambiguës (voir Figure 1.3), bien que certaines relations puissent être modulées (par exemple : "très loin de", "tout à fait à droite, mais un peu au-dessus", etc.), comme cela est fait dans [74] (voir Figure 1.4). Pour des descriptions précises de scènes complexes, on cherchera plutôt à avoir des descripteurs dont l'évaluation est continue, donc quantitative. Un aperçu des techniques de modélisation et d'évaluation des relations spatiales est donné dans la Section 1.1.4.2, ce sujet étant détaillé dans le Chapitre 2 de ce manuscrit.

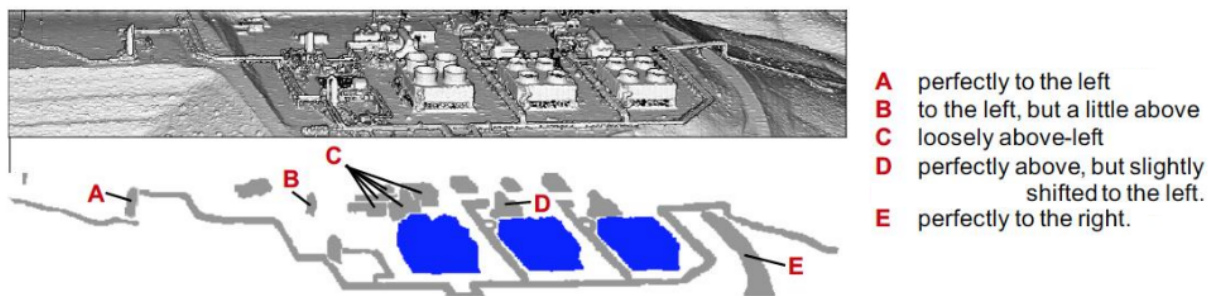


FIGURE 1.4 – Exemple de description de scène par des relations spatiales valuées, pour plusieurs objets ou groupes d'objets par rapport à une même référence (en bleu) (source : [74]).

### Configurations spatiales de scènes ou d'objets

La configuration spatiale peut aussi être décrite en langage naturel par un certain nombre de mots plus ou moins spécialisés. Décrivant un objet ou une scène, elle peut se faire par un adjectif ou par un nom, en utilisant pour cela la préposition "en".

Pour un objet, la forme est une des principales caractéristiques, et est souvent suffisante pour le reconnaître. Par exemple, une silhouette suffit pour détecter la présence d'une personne voire pour la reconnaître. C'est pourquoi il s'agit d'une caractéristique exploitée depuis longtemps en informatique, où la reconnaissance d'objets est souvent assimilée à la "reconnaissance de formes". Cependant, elle ne prend pas du tout en compte la structure de l'objet, c'est-à-dire sa décomposition en parties, ni la configuration de ces parties. Ainsi, dans certains cas la forme peut être suffisante pour reconnaître un objet, mais dans d'autres la configuration spatiale peut avoir un intérêt. Du fait de leur grande variété (presque autant que d'objets), il n'existe pas de taxonomie des formes, excepté pour les formes géométriques, et on préfère souvent décrire une forme par un exemple d'objet qui a une forme similaire ("en forme de ...").

En ce qui concerne la description de scènes, la description la plus pertinente est sans doute celle de sa composition, associée à sa configuration spatiale. Comme pour les formes, il y a une grande diversité de configurations spatiales possibles. Pourtant, le langage contient très peu de termes dédiés à leur description, utilisant plutôt des termes correspondant à des formes, notamment géométriques (en cercle, en étoile...), de lettres de l'alphabet (en L, en U...) ou d'objets avec la même configuration (en peigne, en épi...). En fait, l'organisation des objets d'une scène réelle étant rarement ordonnée, et l'organisation des parties d'un objet n'étant pas toujours rigide, on ne dispose pas de terme adapté pour chaque configuration. Les seuls termes que nous avons identifiés comme véritablement dédiés à cela sont ceux issus du vocabulaire militaire (où l'ordre a son importance...) de configurations "en

bataille" (des "véhicules rangés en bataille") et "en quinconce", désignant à l'origine une formation de la légion romaine. D'autres formations similaires désignent aussi des configurations spatiales, mais elles sont issues de mots désignant autre chose à l'origine et sont peu employées : outre les classiques lignes, colonnes ou carrés, on peut citer les formations triparties (sur trois lignes en quinconce), en éperon (similaire au chevron), en tenaille, etc. En revanche la fameuse "tortue", que nous connaissons bien grâce à Astérix, n'est pas une configuration de bataillons mais une posture des troupes au sein d'un bataillon.

Nous nous proposons ici de donner quelques exemples de configurations spatiales caractéristiques en 2D et en 3D. Cependant, il s'agit plutôt de configurations faisant référence à des formes, puisque comme expliqué précédemment les configurations sont rarement ordonnées. Dans chacun des cas les objets peuvent être disposés sur des points d'intérêt (coins, bords, extrémités) ou répartis sur l'ensemble de la configuration. De plus, lorsque les objets sont dans la même direction et qu'ils ont une forme adaptée (i.e., allongée), ou un avant et un arrière, plusieurs variantes sont possibles : en les disposant soit les uns derrière les autres ("à la queue"), soit en échelons, soit perpendiculaires à la direction, tout en restant parallèles dans chacun des cas. D'autres exemples incluant des configurations indéfinies sont donnés dans les cas d'usage de nos travaux (cf. Section 1.1.3.1).

En 2D, on peut trouver les configurations suivantes :

- des configurations "pures" : en quinconce, en épi, en peigne/bataille... ;
- des formes basiques : en lignes/colonnes, en L, en T, en U, en V (chevron ou éperon), en I... ;
- des formes géométriques plus ou moins complexes : en triangle, carré, cercle, ellipse/ovale, trapèze, étoile, spirale... ;
- des formes d'objets simples : épi, peigne, cacahuète, bouteille, fleur, papillon... ;
- des motifs divers, des combinaisons ou répétitions (type fractales) de formes basiques ;

avec par exemple :

- les formations d'infanterie, comme celles des légionnaires romains : en ligne, en colonne, en carré, en éperon (V ou chevron), en tenaille... ;
- la formation en V utilisée par les oiseaux migrateurs, ou par les patrouilles d'avions pour les défilés aériens ;
- la disposition des tables pour une réception : en L, en T, en U... ;
- des paysages aériens particuliers : habitations en peigne le long des routes, places de parkings en épi ou en bataille, jardins aménagés avec différents motifs... ;
- les configurations du théorème de Thales : triangle ou papillon.

Le passage à la 3D donne alors les configurations suivantes :

- des formes géométriques plus ou moins complexes : cube, parallélépipède, cylindre, tétraèdre, cône, pyramide, sphère, tore (anneau), hélice (forme hélicoïdale/en colimaçon) ;
- des configurations 2D répétées suivant la troisième dimension, ou par rotation autour d'un axe ;
- des motifs divers, des combinaisons ou répétitions (type fractales) de formes basiques ;

avec par exemple :

- certaines formations utilisées par les patrouilles aériennes, ou par les essaims de drones ;
- des éléments de végétation répétant des motifs particuliers : fleurs, arbres (branches d'un sapin), fruits (parties d'une pomme de pin), légumes (tomates sur une grappe)... ;
- des éléments de l'infiniment petit ou de l'infiniment grand : ADN en hélice, molécules avec diverses configurations, électrons et planètes en cercles concentriques, galaxies en spirales... ;
- la configuration de Möbius (couple de tétraèdres mutuellement inscrits).

Un aperçu des techniques de modélisation des configurations spatiales est donné dans la Section 1.1.4.3, ce sujet étant détaillé dans le Chapitre 3 de ce manuscrit. L'ajout de la dimension temporelle conduit alors à des configurations spatio-temporelles, générant encore plus de possibilités (voir les cas d'usage dans la Section 1.1.3.1).

### 1.1.2.4 Prise en compte du point de vue et de la pose des objets

Deux paramètres importants à prendre en compte pour décrire la configuration spatiale d'une scène sont le point de vue et la pose des objets, qui dépend aussi de leur nature.

#### Prise en compte du point de vue, invariance et équivariance

Un point important à noter est que les images réelles sont en général des vues 2D de scènes 3D, donc dépendantes du point de vue de l'observation, c'est-à-dire d'un angle de vue (cf. Figure 1.5) et d'une distance à la scène. Lorsque ce point de vue est modifié, certaines déformations de l'image apparaissent, qui peuvent affecter la façon dont la scène est perçue, et en particulier sa configuration spatiale (voir par exemple la Figure 1.9). De petites variations peuvent avoir des effets limités, tout comme certaines variations particulières : par exemple, une rotation de la caméra autour de son axe de visée engendrera une simple rotation de l'image, une variation de distance engendrera un simple changement d'échelle, tandis qu'une translation de la caméra, perpendiculairement à sa visée, se traduira en une translation dans l'image, permettant de traiter des transformations linéaires. En revanche d'autres variations peuvent avoir des impacts beaucoup plus importants, nécessitant de connaître le mouvement de la caméra pour les estimer et comparer des scènes entre elles, ou de bien connaître les objets observés et d'identifier leur pose pour les retrouver et ensuite déterminer les changements de configuration. Par exemple, une personne vue de dessus n'aura pas du tout la même configuration que vue de face, et des objets bien séparés pourront se superposer si on les voit depuis un autre angle de vue, générant ainsi des occultations.

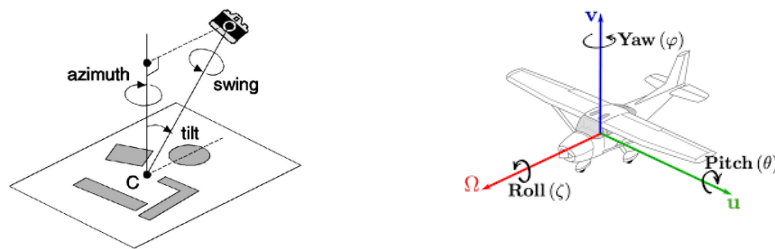


FIGURE 1.5 – Représentation des angles intervenant dans la modélisation du point de vue d'acquisition d'une scène, et des angles définissant l'orientation d'un appareil. Le point de vue angulaire peut être caractérisé par trois angles par rapport au capteur, lorsqu'il est orienté vers la scène : l'azimut, l'inclinaison (*tilt*) ou son complémentaire l'élévation, et le pivotement (*swing*). Ils sont différents des angles de roulis (*roll*), tangage (*pitch*) et lacet (*yaw*), qui définissent l'orientation d'un appareil dans un repère lié à celui-ci et non par rapport à une cible.

Ainsi, la prise en compte de ces variations est essentielle pour comparer correctement deux scènes acquises avec des points de vue différents, et il s'agit d'un critère majeur pour concevoir et évaluer les approches. Le but est de s'affranchir de ces variations, de façon à s'intéresser uniquement aux changements affectant le contenu de la scène. On parle alors d'invariance lorsque le résultat obtenu est le même quelle que soit la variation, et d'équivariance lorsque le résultat évolue de la même manière que la variation. Par exemple, si l'on cherche à prédire la validité de la relation "à droite de" entre deux objets, on s'attend à ce que le score de confiance ne change pas lorsqu'on applique une translation à l'ensemble de la scène (i.e., qu'il soit invariant à une telle translation), mais qu'il évolue dans le même sens lorsqu'on éloigne ou rapproche les deux objets (i.e., qu'il soit équivariant à une telle translation).

Par ailleurs, il peut aussi être important d'avoir des descripteurs robustes à d'autres variations des conditions d'acquisition ou de visualisation de la scène, comme des variations de luminosité, de contraste, de bruit, ou encore la présence d'artefacts ou de dégradations dûs au capteur, au stockage ou à la transmission des données.

## Prise en compte de la nature et de la pose des objets

On a vu qu'il y avait un intérêt à détecter la pose des objets pour déterminer les variations de point de vue d'une acquisition non contrôlée. Pour ce qui est des relations spatiales, notamment directionnelles, connaître la nature et la pose des objets est même essentiel pour bon nombre d'objets orientés ou avec des formes particulières. On appelle objet orienté un objet dont les côtés (ou faces) peuvent être caractérisés par des attributs directionnels de manière absolue, identique quel que soit le référentiel. Concrètement, un objet orienté peut avoir un côté gauche et un côté droit, un dessus et un dessous, un avant et un arrière. Dans ce cas les relations spatiales peuvent être données soit par rapport à ces définitions soit sans en tenir compte. Par exemple, les objets visibles plus à droite que lui dans l'image ne sont pas forcément ceux sur son côté droit, ce qui peut être complexe à modéliser.

Il convient alors de distinguer la perception de la relation spatiale naïve (ou brute), c'est-à-dire sans tenir compte des caractéristiques des objets, et la perception réfléchie (ou inférée), en considérant la nature des objets et leur pose, et de préciser si le type de relation considérée suivant le cas. Ainsi, sur l'exemple de la Figure 1.2, le chat est à droite du chien si l'on ne tient pas compte de leur nature, mais sinon il est en face ! De la même façon, certaines relations topologiques peuvent aussi être interprétées différemment lorsque l'on ajoute la connaissance de l'objet et du point de vue de la scène, plutôt qu'en ne considérant que la perception naïve en 2D, comme c'est le cas pour la relation "sur" par exemple. Le jeu de données *SpatialSense* [119] introduit dans la Section 1.1.3.3 et détaillé dans la Section 1.2.1.3 est un bon exemple de jeu de données contenant des annotations tenant compte de la nature des objets.

### 1.1.3 Applications et jeux de données

La prise en compte de la configuration spatiale peut être très utile pour de nombreuses tâches applicatives et de nombreux cas d'usage en vision par ordinateur. Dans un premier temps, nous cherchons à lister les types de scènes ou d'objets pour lesquels il est pertinent d'étudier la configuration spatiale, afin de les reconnaître ou de les suivre. Puis nous détaillons les tâches applicatives et les cas d'usage où elle a un intérêt. Enfin, nous donnons des exemples de jeux de données de la littérature considérant des tâches qui s'intéressent aux configurations spatiales.

#### 1.1.3.1 Types de configurations d'intérêt

La configuration spatiale peut intervenir à deux niveaux dans le cadre de la comparaison de scènes ou d'objets : pour appairer les composantes entre elles et pour les comparer une fois appariées, d'autres caractéristiques pouvant être exploitées pour chacune des deux étapes. Pour la reconnaissance, l'ajout de la dimension temporelle permet en plus de prendre en compte l'évolution particulière de la configuration, produisant des configurations spatio-temporelles et permettant d'aborder d'autres cas d'usage à partir de séquences d'images. On distinguera alors ces configurations de celles où l'évolution n'apporte pas d'information supplémentaire, où l'on peut donc traiter chaque image indépendamment.

#### Pour l'appariement

Pour l'appariement en lui-même, la configuration spatiale est utile pour de nombreux cas d'usages, en tant que première étape avant d'autres traitements qui ne font pas forcément appel à elle. Par définition, l'appariement basé sur la configuration spatiale est possible pour des scènes et objets dont la configuration est proche. Cela s'applique donc à des scènes ou objets rigides (paysage aérien, ensemble de bâtiments, véhicule...), peu déformables (visage, personne, animal, végétal, vêtement, groupes d'objets organisés...), ou déformables mais ayant peu bougé, ce qui est le cas entre des images proches dans une séquence vidéo typiquement.

### **Pour la reconnaissance de configuration spatiale sans mouvement**

Les cas d'usages identifiés pour la comparaison d'objets à l'aide de la configuration spatiale couvrent à la fois la reconnaissance de scènes en tant qu'ensemble d'objets ou de points, ou d'objet en tant qu'ensemble de parties ou de points d'intérêt. Dans une scène, les cas typiques sont ceux de groupes d'objets organisés, comme un convoi, une flotte, une patrouille aérienne... (e.g., encerclement, vol en formation), ou de scènes aériennes particulières (paysage, complexe industriel, port, ville, village, cadastre, ou encore constellation en regardant vers le ciel), par opposition à la détection d'un tel groupe d'objets sans prendre en compte la configuration, où la co-occurrence serait alors suffisante. Pour un objet, la configuration spatiale de parties ou points d'intérêt peut permettre de reconnaître un modèle de véhicule, une espèce animale ou végétale, un vêtement, un symbole, ou encore la posture d'une personne ou une expression de visage. En revanche, bien qu'elle puisse être utile pour apparier des visages ou des personnes, elle ne l'est pas vraiment pour les distinguer. Enfin, lorsque l'agencement est moins important, la forme globale peut être suffisante, pour reconnaître un type de véhicule ou certains ensembles de bâtiments par exemple.

### **Pour la reconnaissance de configuration spatio-temporelle**

Lorsque l'on cherche à reconnaître une configuration spatio-temporelle, la configuration spatiale n'a un intérêt que si elle varie dans le temps. Elle passe alors par l'étude des trajectoires relatives des objets, c'est-à-dire de l'évolution des positions relatives. Si la configuration est stable, alors l'analyse des trajectoires absolues est suffisante. C'est le cas par exemple pour un groupe, un convoi, une flotte, une patrouille aérienne organisés qui auraient la même trajectoire. Il existe cependant quelques cas d'usage où l'on peut chercher à reconnaître une configuration spatio-temporelle particulière, notamment pour un groupe d'objets avec un déplacement différent mais coordonné et caractéristique. La configuration peut alors être plus ou moins complexe :

- simple : attroupement/encerclement, élément qui se déplace à contre-sens des autres, mouvement des parties du corps d'une personne, expression de visage ;
- complexe : danse, chorégraphie, stratégie d'attaque (sport/bataille), convoi, flotte, patrouille aérienne organisé(e) avec des trajectoires différentes.

Dans une moindre mesure, la configuration spatio-temporelle peut avoir un intérêt pour reconnaître un déplacement coordonné mais "flou", c'est-à-dire avec une tendance et des variations, comme un déplacement dans la même direction à long-terme, un mouvement de foule ou une stratégie d'attaque.

#### **1.1.3.2 Tâches applicatives et cas d'usage**

La modélisation de la configuration spatiale peut être très utile pour de nombreuses tâches de vision par ordinateur, des tâches "basiques" comme la simple comparaison ou description, plus évoluées comme la reconnaissance ou la recherche d'images similaires, à des tâches appliquées comme le suivi d'objets, l'aide à la navigation, etc. En voici une liste non exhaustive :

- la description de scène en langage naturel, le VQA, le *Visual Reasoning*, avec comme cas d'usage les IHM pour la robotique ou les systèmes d'assistance vocale (audio-description, aide à la navigation) par exemple ;
- la reconnaissance d'objets ou de scènes ;
- la recherche d'objets ou de scènes similaires dans une base (CBIR, ou plus simplement "retrieval") ;
- la détection de changements, dans des séquences typiquement ;
- la détection ou la prédiction de situation d'intérêt ou à risque, c'est-à-dire de situations connues, comme une stratégie d'attaque en sport collectif ou sur un champ de bataille ;
- la détection ou la prédiction d'anomalie, c'est-à-dire de situations ou d'éléments inattendus par rapport à ce qui est connu ;



- la reconnaissance ou la prédiction de l'évolution d'une situation, pour la vidéo-surveillance ou la navigation autonome par exemple.

Les cas d'usage couvrent de nombreux domaines d'application :

- en sécurité et Défense : vidéosurveillance, analyse d'images et vidéos aériennes/satellites ;
- en médecine : aide au diagnostic en imagerie 2D/3D (OCT, radiographie, scanner, TEP, IRM, échographie...);
- en robotique : localisation dans l'environnement et dialogue (IHM), localisation basée vision en navigation aérienne ;
- pour le transport : assistance à la navigation, navigation autonome ;
- dans la vie quotidienne : description de scène pour les mal-voyants ou les personnes à mobilité réduite, reconnaissance de scène à risque (personne au sol par exemple)... ;
- en botanique ou zoologie : reconnaissance (taxinomie et systématique) des espèces végétales ou animales ;
- en architecture, paysagisme, urbanisme, art, design : recherche de configurations similaires ;
- dans le sport : analyse de stratégies.

Le cas d'usage qui nous intéresse principalement ici est celui général de la reconnaissance de scènes ou d'objets (ou plutôt la ré-identification), appliqué à la vidéo drone, cas d'usage détaillé dans le Chapitre 6 de ce manuscrit sur la ré-identification de personnes.

### 1.1.3.3 Jeux de données

Peu de jeux de données sont véritablement dédiés à l'analyse de configurations spatiales, que ce soit sur des images naturelles ou synthétiques, pour des configurations de deux objets ou plus. Il en existe cependant quelques-uns, notamment pour des tâches de raisonnement spatial (*Spatial Reasoning*, VQA), où il s'agit de répondre à des questions précises sur les relations spatiales entre certains objets, ou pour des tâches de *captioning*, où il s'agit de décrire le contenu de la scène en langage naturel, notamment en termes de relations entre les objets. Ces descriptions peuvent alors être utilisées dans une interface homme-machine, ou pour évaluer la capacité d'une intelligence artificielle à gérer ce type de concepts.

Concernant le raisonnement visuel et le VQA, on peut par exemple citer les jeux de données 2D simulées *SHAPES* [2]<sup>2</sup>, *ShapeWorld* [60]<sup>3</sup> ou *NLVR* [108]<sup>4</sup>, contenant tous les trois des formes géométriques simples (cercles, polygones, croix), avec des couleurs, des tailles, des orientations différentes (pour *ShapeWorld* seulement) et surtout, pour ce qui nous intéresse ici, des positions différentes (pour tous). Ces données sont générés en spécifiant des contraintes de localisation des objets, par exemple en évitant le recouvrement ou les relations directionnelles ambiguës, ce qui permet d'obtenir directement des annotations. Le jeu de données CLEVR [54]<sup>5</sup> contient le même type de contenu mais avec des images naturelles et du contenu 3D, puisqu'il s'agit de photographies de scènes contenant des objets de formes géométriques simples (cubes, sphères, cylindres), de taille, de couleur et d'aspect variables, positionnés de façon variable dans la scène. Le contenu de la scène est contrôlé mais nécessite alors un travail d'annotation manuel. Des exemples de ces jeux de données sont donnés dans la Figure 1.6.

Dans la tâche de *captioning*, le but est de décrire le contenu de la scène en langage naturel, soit avec une phrase succincte soit de façon exhaustive en décrivant l'ensemble des objets présents (noms), leurs relations (prépositions ou verbes), ainsi que les actions qu'ils exécutent (verbes). Une tâche qui en dérive concerne spécifiquement la détection de relations visuelles (*visual relationship detection*), dont le but est de détecter des triplets (sujet, relation, objet), avec différents types de relations possibles. Plusieurs jeux de données d'images naturelles ont été conçus et annotés pour cette tâche, comme *Visual*

2. <https://github.com/ronghanghu/n2nmn#train-and-evaluate-on-the-shapes-dataset>

3. <https://github.com/AlexKuhnl/ShapeWorld>

4. <https://lil.nlp.cornell.edu/nlvr/>

5. <https://cs.stanford.edu/people/jcjohns/clevr/>

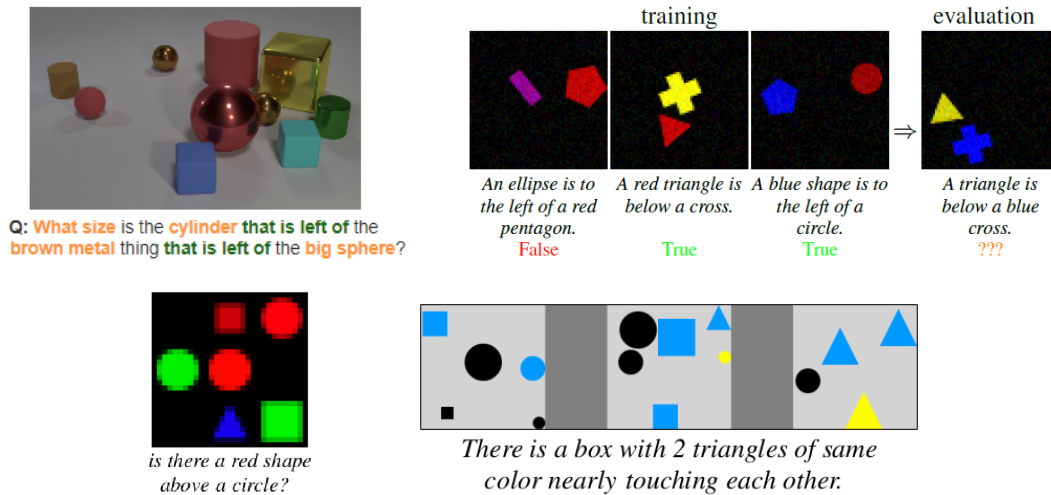


FIGURE 1.6 – Exemples de données dédiées au raisonnement spatial, issues des jeux de données CLEVR [54], ShapeWorld [60], SHAPES [2] et NLVR [108].

*Relationship Dataset (VRD)* [68]<sup>6</sup>, *Visual Genome* [57]<sup>7</sup>, ou encore *Open Images* [62]<sup>8</sup>, qui contiennent tous des objets, relations et actions variés. D'autres jeux de données similaires ont été conçus ou annotés spécifiquement pour apprendre à reconnaître directement des triplets (sujet, relation spatiale, objet), comme *SUN09* [14] (avec les annotations issues de [63]<sup>9</sup> puis de [69]<sup>10</sup>) ou *SpatialSense* [119]<sup>11</sup>. Ce dernier a été utilisé dans nos travaux, comme détaillé dans la Section 1.2.1.3. Enfin, le jeu de données *Rel3D* [42]<sup>12</sup> contient pour sa part des images simulées de scènes 3D contenant chacune deux objets avec une relation définie. Il est destiné à l'apprentissage de relations prenant en compte la pose des objets, puisqu'il contient des couples de configurations similaires où une relation donnée est soit vraie soit fausse en fonction de la pose. Des exemples de ces jeux de données sont présentés dans la Figure 1.7.

Par ailleurs, on peut aussi citer les travaux de [111] sur la recherche de configurations 2D similaires, utilisant pour cela une image simulée contenant plusieurs couples d'objets similaires. Cette image est reproduite dans la Figure 1.8. Les configurations présentes ont été choisies afin d'imiter des configurations qui peuvent être trouvées dans des images satellitaires, comme dans les exemples de la Figure 1.1. De telles images peuvent alors être utilisées pour rechercher des configurations spatiales comme cela est fait dans [111], mais elles nécessiteraient des annotations dédiées pour une évaluation quantitative.

#### 1.1.4 Axes de recherche de la littérature

Nous donnons ici un aperçu des techniques de modélisation des configurations spatiales, entre deux objets puis pour une scène composée de plusieurs objets. Ces sujets sont détaillés dans les Chapitres 2 et 3 de ce manuscrit, dans les Sections 2.1 et 3.1 respectivement.

6. <https://cs.stanford.edu/people/ranjaykrishna/vrd/>

7. <https://visualgenome.org/>

8. <https://opensource.google/projects/open-images-dataset>

9. <https://cs.stanford.edu/~taranlan/>

10. <https://www.mpi-inf.mpg.de/departments/computer-vision-and-machine-learning/research/vision-and-language/learning-spatial-relations/>

11. <https://github.com/princeton-vl/SpatialSense>

12. <https://github.com/princeton-vl/Rel3D>



### 1.1.4.1 Généralités

Trois stratégies sont souvent considérées pour traiter les objets dans les différents modèles présentés : par leur barycentre, par leur rectangle englobant ou par leur masque de segmentation (ou contour). Cependant, le masque et le barycentre supposent qu'une segmentation a été effectuée, ce qui peut être coûteux et générer des erreurs. Le rectangle englobant est un peu moins exigeant quant à lui en supposant seulement qu'une détection a été effectuée. Par ailleurs, si l'on a besoin de prendre en compte la pose et la nature des objets, il faudrait en plus extraire cette pose et cette sémantique, en utilisant des traitements adaptés. Enfin, la décomposition des objets pourrait aussi être utile dans certains cas, pour décrire des postures ou des comportements par exemple.

### 1.1.4.2 Descripteurs de configuration spatiale entre deux objets

Plusieurs études ont été menées sur l'analyse des relations spatiales dans différents domaines d'application de la reconnaissance de formes et de la vision par ordinateur, avec l'objectif commun de décrire la configuration spatiale des objets dans les images. Les premiers travaux formalisant ce type d'information entre deux objets sont ceux de Freeman, qui en 1975 proposa une catégorisation en 13 relations spatiales qualitatives [39]. D'autres catégorisations similaires ont été proposées ensuite, comme les RCC [95, 23] qui font toujours référence sur cet aspect (cf. Figure 2.2), ainsi que plusieurs solutions utilisant des projections des objets sur une seule dimension, comme celle de [56] basée sur les intervalles temporels d'Allen [1] (cf. Figure 2.1). Ces relations de nature qualitative peuvent être directionnelles ("*à gauche de*", "*au-dessus de*", etc.), topologiques ("*à l'intérieur de*", "*à l'extérieur de*", "*tangent à*", etc.) ou distancielles ("*proche de*", "*loin de*"). Une telle approche qualitative a l'avantage de fournir directement une description en langage naturel, mais sa nature "tout ou rien" n'est manifestement pas adaptée pour décrire des relations plus complexes ou ambiguës.

C'est pourquoi Freeman a également suggéré d'utiliser des relations dites "floues" [39]. Dans cette approche, une relation spatiale qualitative telle que "*à gauche de*" est considérée, et une évaluation quantitative de cette relation est obtenue pour deux objets donnés. La mesure peut être basée soit sur l'orientation, la distance, le recouvrement ou les dimensions relatives des objets. Cependant, à cette époque les capacités des ordinateurs n'ont pas permis de modéliser efficacement ces concepts spatiaux fondamentaux. De nombreux auteurs ont alors assimilé les objets  $2D$  à des entités très élémentaires telles qu'un point (centroïde) ou un rectangle (englobant). La procédure est commode, mais ne peut pas fournir une modélisation satisfaisante, comme l'a souligné Rosenfeld en 1982 [96], notamment pour des configurations complexes (voir Figure 1.3). Ce n'est qu'après plusieurs années que les relations floues ont pu réellement être exploitées, par exemple avec les paysages flous [3, 4].

Une approche parallèle consiste à utiliser les mesures quantitatives elles-mêmes comme descripteurs de la configuration spatiale, afin d'avoir une description plus fine. On parle alors de descripteurs de position relative lorsque le but est de décrire complètement la configuration, en utilisant une combinaison de plusieurs mesures typiquement. Un exemple célèbre de descripteur de position relative est l'histogramme de forces [79] (cf. Sections 2.1.1.2 et 2.2.2.1), qui combine l'aspect directionnel et des mesures de distance. De cette manière, la position relative d'un objet par rapport à un autre peut avoir une représentation qui lui est propre, ce qui permet de l'utiliser comme descripteur dans un système de reconnaissance par exemple. Il est aussi possible d'en déduire des évaluations de différentes relations spatiales en langage naturel, comme cela est fait dans [79, 74] (voir Figures 1.4 et 2.7).

Ainsi, nous pouvons distinguer deux types de descripteurs de configuration spatiale entre deux objets, avec une dualité forte :

1. les descripteurs de relations spatiales, décrivant la configuration selon une ou plusieurs relations spatiales qualitatives, chaque relation pouvant être évaluée par des mesures quantitatives spécifiques, en tant que relations "floues";

2. les descripteurs de position relative, décrivant complètement la configuration par des mesures quantitatives, souvent selon différentes directions, et qui peuvent être utilisés pour évaluer différentes relations spatiales.

#### 1.1.4.3 Description de configuration spatiale d'une scène ou d'un objet

Les descripteurs dédiés à la forme sont parmi les premiers qui ont été développés, notamment pour la reconnaissance d'objets, qui a d'ailleurs pris la dénomination de "reconnaissance de formes". En revanche, à notre connaissance, très peu de travaux ont abordé la description de la configuration spatiale. De la même façon que pour les descripteurs de relations spatiales, on pourrait avoir des descripteurs dédiés à l'évaluation de certaines configurations spatiales, par exemple pour évaluer la présence d'une configuration "en V" ou "en quinconce". Cependant, les descriptions de configurations en langage naturel ne permettent de décrire que certaines configurations remarquables et simples, alors que la plupart des configurations sont désordonnées et complexes.

Une solution est de représenter la scène sous la forme d'un graphe relationnel dont les nœuds sont les constituants de la scène et les arêtes les relations entre ces constituants. Les graphes sont en effet les structures naturelles pour ce type de modélisation, mais ils ont le défaut d'être difficilement comparables. Une autre option couramment utilisée est de trier les objets en les projetant sur un axe afin de pouvoir comparer plus facilement les représentations. C'est par exemple l'approche des *2D-strings* [13], qui utilisent une projection suivant l'axe des abscisses et l'axe des ordonnées, ou des *k-formules* dans [110], qui utilisent un nombre plus élevé de directions et décrivent la scène suivant chacune de ces directions. Cependant, ces représentations sont très sensibles à des variations individuelles des objets constituant la scène, qu'il n'est pas aisé de localiser. De plus, les descripteurs dédiés aux relations spatiales nécessitent d'abord de détecter voire de segmenter les objets d'intérêt, ces descripteurs étant calculés sur des images binaires en général. Ainsi, il n'y a eu que très peu d'utilisations de ces descripteurs jusqu'à maintenant, et leur intégration dans un modèle de scène reste un défi majeur.

#### 1.1.4.4 Prise en compte du point de vue

Comme introduit dans la Section 1.1.2.4, un problème majeur est la grande sensibilité de l'image au point de vue 3D, ce qui nécessite de prendre en compte la pose des objets, donc leurs positions, leurs orientations et celles du capteur lors de l'acquisition de l'image. Plusieurs approches sont possibles pour cela. La première consiste à utiliser des descripteurs invariants à ces changements, ou d'appliquer une normalisation à des descripteurs non invariants lorsque c'est possible, comme avec l'histogramme de forces (cf. Section 2.1.1.3). Une autre approche consiste à recalculer les scènes en amont du calcul des descripteurs, ce qui revient à faire de l'appariement de scènes, mais pour cela il peut justement être utile de considérer la configuration spatiale. Une piste intéressante serait alors de réaliser conjointement le recalage et la comparaison.

D'autres approches cherchent à modéliser la structure des objets avec des modèles plus explicables et robustes aux variations de point de vue. Il s'agit notamment des réseaux à capsules (*capsule networks*) [48] dont l'idée est de vérifier que les parties sont présentes avec les mêmes paramètres, par exemple en termes de pose. Cependant les implémentations actuelles de ce concept, comme celle de [98], n'ont pas permis d'atteindre cet objectif pour le moment [84, 38], malgré des avancées intéressantes dans la modélisation [33, 94].

### 1.1.5 Défis et positionnement des travaux

Plusieurs défis majeurs restent à relever afin de pouvoir exploiter correctement la configuration spatiale dans des tâches de compréhension de scène, notamment :

**1. la modélisation de relations et de configurations spatiales pour des configurations complexes ou ambiguës :**

Des descripteurs existent déjà pour décrire la position relative de deux objets, dans le cas général et pour certaines relations particulières (cf. Section 1.1.4), mais ils ne permettent pas de décrire précisément toutes les situations et en particulier les configurations complexes. Le développement de descripteurs plus précis et facilement manipulables reste donc un sujet d'étude important. Dans nos travaux nous proposons un nouveau descripteur de position relative dédié aux configurations complexes entre deux objets, qui est détaillé dans le Chapitre 2 de ce manuscrit.

**2. le passage de descripteurs de relations spatiales entre deux objets à des descriptions de configurations de scènes comportant plusieurs objets :**

Les descripteurs existants sont destinés à des configurations de deux objets, alors que les scènes ou les objets réels sont constitués d'un nombre important d'objets ou de parties. Quelques modèles ont été envisagés mais aucune étude conséquente sur le sujet n'a encore été menée. Le Chapitre 3 de ce manuscrit est dédié à cela, proposant une solution et donnant des pistes pour une étude plus approfondie.

**3. l'utilisation de descripteurs de configuration spatiale sur des images réelles non segmentées :**

Les descripteurs existants sont calculés sur des formes binaires, ce qui nécessite une étape de segmentation en amont pour les utiliser sur des images réelles, et très peu de travaux ont traité ce problème. Deux pistes sont alors possibles : développer des méthodes de segmentation adaptées aux données dont on souhaite exploiter la configuration, ou développer des méthodes de description sur des images non segmentées, avec une approche de bout-en-bout. Nous avons dû faire face à ce problème dans nos travaux et nous avons proposé deux méthodes de segmentation qui sont détaillées dans la deuxième partie de ce manuscrit (Chapitres 4 et 5).

**4. la prise en compte de la pose des objets et du point de vue dans la scène 2D, avec possibilité d'invariance selon la tâche :**

Afin de pouvoir comparer des scènes acquises dans des conditions différentes, il est nécessaire d'utiliser des approches capables de reconnaître ces variations, de façon à les prendre en compte de façon adaptée à la tâche, c'est-à-dire en les intégrant dans le calcul ou en les supprimant suivant le besoin. Des solutions existent déjà pour cela, avec des descripteurs dédiés et des méthodes de normalisation par exemple, mais elles sont encore peu utilisées du fait de certaines limites ne permettant pas de les utiliser facilement pour des tâches concrètes.

**5. la prise en compte de la pose des objets et du point de vue dans la scène 3D, avec possibilité d'invariance selon la tâche :**

Les images traitées sont en général des vues 2D de scènes 3D, et les mouvements dans la scène 3D génèrent des modifications difficiles à prendre en compte dans la scène 2D, qui requièrent des modèles plus évolués. Or, les descripteurs de configuration spatiale existants sont destinés aux scènes 2D et ne permettent pas de décrire des configurations dans la scène 3D ni de prendre en compte des modifications dans la scène 3D. Des méthodes de détection de la pose des objets pourraient être utilisées pour cela, mais elles n'ont pas encore été intégrées dans la description de configuration spatiale à notre connaissance.

**6. l'intégration des descripteurs dans des chaînes de traitement ou des modèles existants :**

Les descripteurs de configuration spatiale existants ont été proposés dans un cadre théorique en tant qu'outils pouvant être utilisés pour décrire la scène, mais très peu d'exemples d'intégration existent en tant que composante d'une chaîne de traitement destinée à une tâche réelle plus complexe. L'utilisabilité de ces approches pour de telles tâches reste donc à explorer. Dans nos travaux, nous avons proposé une chaîne de traitement globale pour répondre à une tâche de ré-identification. Celle-ci est détaillée dans le Chapitre 6 de ce manuscrit.

## 7. la disponibilité de données pour entraîner et évaluer les approches :

La disponibilité de données annotées est un problème majeur en vision par ordinateur, pour évaluer les algorithmes, mais surtout pour entraîner les modèles usuels basés sur l'apprentissage, d'autant plus pour l'apprentissage profond qui nécessite de grandes quantités de données pour cela. Les approches plus traditionnelles de description par des modèles explicites ne nécessitent pas autant de données que les approches d'apprentissage profond, mais doivent quand même être évaluées, et éventuellement entraînées, sur des jeux de données de taille suffisante. Or, très peu de jeux de données existent pour les tâches basées sur la configuration spatiale, et ceux-ci ne sont pas toujours facilement utilisables pour une autre tâche que celle prévue initialement. La Section 1.1.3.3 de ce chapitre répertorie quelques jeux de données existant dans la littérature, tandis que la Section 1.2 détaille les approches utilisées afin de disposer de données pour nos travaux.

Dans nos travaux, nous avons particulièrement exploré les points 1, 3 et 7, et également traité les points 2 et 6. Les points 4 et 5 ont aussi été abordés mais sans évaluation spécifique.

## 1.2 Génération de données

Une partie importante de nos travaux a été de rechercher et adapter des jeux de données existants, voire de concevoir de nouveaux jeux de données, afin d'évaluer les différentes approches que nous proposons. Nous détaillons cela dans cette section, pour les relations spatiales entre deux objets dans un premier temps, et pour les configurations de scènes ou d'objets composés de plusieurs parties dans un second temps.

### 1.2.1 Données pour les relations spatiales entre deux objets

#### 1.2.1.1 Génération et annotation d'un jeu de données synthétique

Il s'agit d'images binaires contenant deux objets et un arrière-plan uni, qui ont été générées en disposant différentes formes à des positions aléatoires et avec des paramètres d'orientation, d'échelle et de déformation aléatoires. Deux jeux de données ont été produits :

- *2SimpleShapes1* (S1) : 1 000 images de paires d'objets (10 images par paire) de taille  $224 \times 224$ , obtenues à partir de 10 formes différentes représentant des objets basiques (maisons, avions, voitures, etc.) construits à partir de formes géométriques simples (triangles, rectangles, ellipses). Toutes ces paires ont en général des configurations simples puisque tous les objets sont totalement remplis, barycentriques (*i.e.*, construits autour de leur barycentre) et ont des tailles comparables ;
- *2SimpleShapes2* (S2) : 1 280 images de paires d'objets (20 images par paire) de taille  $224 \times 224$ , obtenues à partir de 8 formes géométriques simples (triangles, rectangles, semi-ellipses), petites ou allongées, avec des paramètres variables. Ces paires ont des configurations plus complexes, étant donné qu'elles contiennent des objets de tailles variables et des semi-ellipses, qui ne sont pas barycentriques.

Ces images ont été annotées manuellement pour fournir une relation spatiale pour chaque scène selon quatre classes correspondant aux quatre directions principales ("à droite", "à gauche", "au-dessus", "en-dessous"), avec un consensus de trois experts pour les cas ambigus. Pour chaque image, un objet a été considéré comme le référent et l'autre comme l'argument (cf. Figure 2.9), et une relation spatiale a été choisie, sans tenir compte de leur pose (*i.e.*, en considérant que les objets n'ont pas de côté gauche/droit, ni de dessus/dessous). De plus, nous considérons que la relation opposée est symétrique, ce qui donne potentiellement deux fois plus d'annotations en les inversant.

En raison du caractère aléatoire du processus génératif, les jeux de données contiennent diverses configurations spatiales allant de configurations simples à des configurations plus complexes pouvant

conduire à des situations plus ambiguës. Les images ont également été triées selon la complexité et le niveau d’ambiguïté de la relation spatiale (en quatre niveaux différents, de  $N1$  pour les cas simples à  $N4$  pour les cas ambigus), de manière à évaluer séparément chaque niveau. Les cas vraiment ambigus qui n’étaient pas décidables ( $N4$ ) ont été rejetés des jeux de données dans les expériences. Après cette opération, il reste 866 images dans  $S1$  et 1 127 dans  $S2$ . La composition de chaque jeu de données est indiquée dans le Tableau 1.1, tandis que la Figure 2.13 donne un exemple pour chaque relation et chaque niveau de complexité. On peut remarquer que  $S2$  contient des configurations relativement plus complexes que  $S1$ , ce qui était justement désiré avec l’utilisation d’objets plus variés.

Nous avons choisi d’annoter ces données à la main malgré le mode de génération contrôlé car il s’agit de configurations complexes, où la position du barycentre ne suffit pas pour inférer la relation, comme nous le verrons dans les résultats de classification de relations du Chapitre 2. Afin de faciliter et accélérer la tâche, nous avons conçu une mini-IHM en *python* permettant de présenter automatiquement les images à l’annotateur et de renseigner plus facilement les fichiers d’annotation.

TABLEAU 1.1 – Composition des différents jeux de données en termes de complexité, des relations faciles ( $N1$ ) aux relations totalement ambiguës ( $N4$ , exclues des tests).

	$N1$	$N2$	$N3$	$N4$	Total	$N1$	$N2$	$N3$	$N4$
<i>2SimpleShapes1</i> ( $S1$ )	634	155	77	134	1000	63,4%	15,5%	7,7%	13,4%
<i>2SimpleShapes2</i> ( $S2$ )	787	218	122	153	1280	61,5%	17,0%	9,5%	12,0%
<i>image SIG</i> ( $SIG$ )	126	41	15	21	211	59,7%	23,2%	7,1%	10,0%

### 1.2.1.2 Découpage et annotation d’une image de télédétection

Afin d’évaluer notre approche sur des images réelles, nous avons également utilisé une image optique de télédétection déjà segmentée, en utilisant les différentes couches pour extraire des paires d’objets. Il s’agit de l’image d’une zone urbaine contenant divers éléments : maisons, routes, rivière, champs..., dans différentes couches. Nous avons utilisé les couches pour n’en garder que deux à la fois, de façon à obtenir des configurations entre deux objets (ou ensembles d’objets) de la même classe sémantique. Nous avons ensuite découpé l’image en tuiles de taille  $224 \times 224$  à différentes résolutions pour obtenir des patches au contenu varié et en nombre significatif (211), que nous avons annotés comme précédemment par relation et par niveau de difficulté. Le jeu de données obtenu est appelé *image SIG* ( $SIG$ ). Il est intéressant de noter qu’il contient des formes composées de plusieurs parties qui ne sont pas connexes, ce qui peut entraîner des situations plus complexes (cf. Figure 2.9, 3<sup>e</sup> colonne). Le Tableau 1.1 donne la composition du jeu de données selon la difficulté.

### 1.2.1.3 Segmentation d’un jeu de données réelles annotées

*SpatialSense* [119] est fourni avec des annotations de détections (rectangles englobants) et de relations spatiales (triplets (sujet, relation, objet)), comme illustré avec quelques exemples sur la Figure 1.7c.

Ce jeu de données est composé de 11 569 images de scènes de la vie quotidienne, des animaux, des personnes, mais aussi des "objets" d’arrière-plan (ciel, sol, murs, etc.), provenant de *Flickr* ou *NYU Depth*, avec des tailles variables. Celui-ci a été annoté par production participative (*crowdsourcing*) avec une approche adversaire, en entraînant le système à proposer aux annotateurs des relations difficiles à prédire à l’aide d’indices simples tels que la configuration spatiale  $2D$  ou des a priori sémantiques. Il contient neuf relations spatiales différentes, dont les quatre précédentes, et un total de 17 498 annotations, mais une moitié de correctes et l’autre de fausses pour chaque relation, selon la répartition donnée dans le Tableau 1.2.

Nous avons alors exploité les annotations des objets (étiquettes et boîtes englobantes) pour segmenter les deux objets impliqués dans différentes relations, selon la méthode présentée dans le Chapitre 4.

TABLEAU 1.2 – Composition des annotations du jeu de données *SpatialSense* par relation. Pour chaque relation, il y a autant d’annotations valides et non valides ; les nombres indiqués sont ceux pour l’ensemble du jeu de données.

on	behind	in front of	next to	under	in	above	to the left of	to the right of	TOTAL
4866	2958	2174	1562	1490	1358	1262	1026	802	17498

Lorsque le score de confiance de la segmentation n’était pas suffisant, nous avons préféré prendre le rectangle englobant entier. Pour nos expérimentations nous avons uniquement considéré les quatre relations cardinales, ce qui donne 2 290 annotations, et la relation "in" qui en contient 679, ce qui donne un total de 2 969 annotations.

Par ailleurs, il est à noter que les relations de *SpatialSense* sont conçues pour des modèles prenant en compte les 3 dimensions de la scène, en les donnant du point de vue de l’objet référent (cf. Section 1.1.2.4). Par exemple, un homme vu de face conduira à des relations inversées par rapport aux annotations 2D. De ce fait, ce jeu de données est beaucoup plus difficile à appréhender et nécessite des méthodes adaptées.

## 1.2.2 Données pour la configuration spatiale d’une scène ou d’un objet

### 1.2.2.1 Génération de scènes simulées avec *Blender*

*Blender* est un outil de simulation qui permet de générer des scènes 3D contenant divers objets et décors, pouvant être utilisé pour produire des scènes variées et réalistes. La simulation permet de contrôler totalement le contenu de la scène et les conditions de prises de vue, ce qui évite les aléas de l’acquisition et le lourd travail d’annotation (suivant l’application). Nous avons utilisé cet outil pour générer des scènes simples uniquement, en utilisant comme objets des formes géométriques comme des parallélépipèdes ou des cylindres. Plusieurs séquences ont été générées avec différentes variations, comme une rotation autour de la scène ou la translation d’un des objets. Des extraits d’une séquence combinant ces deux variations sont présentées dans la Figure 1.9. Elle a notamment été utilisée pour l’étude du suivi de configurations spatio-temporelles (cf. Section 3.3).

### 1.2.2.2 Extraction des objets dans une scène

Afin d’évaluer les descripteurs de configuration spatiale sur des scènes comportant plusieurs objets, nous avons utilisé la séquence *CamSeq01* introduite dans [37]<sup>13</sup>. Il s’agit d’une séquence d’images acquises dans un environnement urbain qui ont été segmentées avec les principaux objets présents, pour le cas d’usage de la navigation autonome. La séquence originale contient 101 images et jusqu’à 32 classes par image : éléments de "décor" de la route (route, trottoir, signalisation) ou de l’environnement (ciel, immeubles, lampadaires...), qui sont immobiles s’il n’y a pas de mouvement du véhicule, et d’autres usagers mobiles (véhicules, piétons, cyclistes...). Nous avons sélectionné une sous-séquence de 11 images et utilisé la segmentation fournie pour en extraire 12 objets (parfois de la même classe). La Figure 1.10 contient 3 images extraites de cette séquence avec leurs segmentations.

Ce jeu de données n’a pas été utilisé dans nos travaux mais dans le cadre d’un projet d’étudiants. Il pourra néanmoins être utilisé ultérieurement comme jeu de test pour la comparaison de configurations spatiales. D’autres jeux de données similaires existent, comme *CityScapes* ou *SYNTHIA* qui sont couramment utilisés pour ce cas d’usage.

13. <http://mi.eng.cam.ac.uk/research/projects/VideoRec/CamSeq01/>



### 1.2.2.3 Décomposition d'objets en parties dans des séquences

Le cas d'usage principal intéressant l'entreprise Magellium étant celui de la reconnaissance d'objets dans des vidéos aériennes, et en particulier la ré-identification de personnes, nous avons également cherché à évaluer nos descripteurs sur de telles données (cf. Chapitre 6). Pour cela, nous avons utilisé le jeu de données *AeroScapes* [87]<sup>14</sup>. Celui-ci contient plusieurs séquences acquises par un mini-drone à basse altitude, qui est téléguidé pour suivre diverses cibles, notamment en vue oblique : piéton, véhicule, vélo... Il est fourni avec des annotations de segmentations (au niveau pixels) des principaux objets de la scène : éléments de décor (ciel, végétation, routes, signalisation) et éléments mobiles (véhicules, personnes). Pour notre cas d'usage nous avons extrait les personnes dans trois séquences différentes, illustrées sur la Figure 1.11 : les séquences 000001 et 000002 où le drone survole une zone avec une personne en train de marcher tout en tenant un petit objet (qui semble être la manette du drone), à distance pour la séquence 000001 et plus près pour la séquence 000002, et la séquence 040000 où il suit une personne en train de courir sur un chemin, à distance moyenne.

Nous avons choisi ce jeu de données car la résolution des images était suffisante pour traiter le cas d'usage de la reconnaissance de personnes en se basant sur leur allure et leur tenue vestimentaire, en décomposant en parties et en utilisant la configuration spatiale pour apparier les parties. Pour cela, nous avons extrait les portions d'images correspondant aux personnes, soit en gardant uniquement les pixels correspondants, soit tout le rectangle englobant (comme s'il s'agissait d'une détection). Pour exploiter la configuration des parties, nous avons alors cherché à extraire des parties pertinentes et stables au cours de la séquence. Pour cela nous avons utilisé une segmentation basée sur la couleur et séparé les composantes connexes, selon la méthode décrite dans le Chapitre 5. Des résultats sur plusieurs images des séquences 000002 et 040000 sont donnés sur les Figures 1.12 et 1.13.

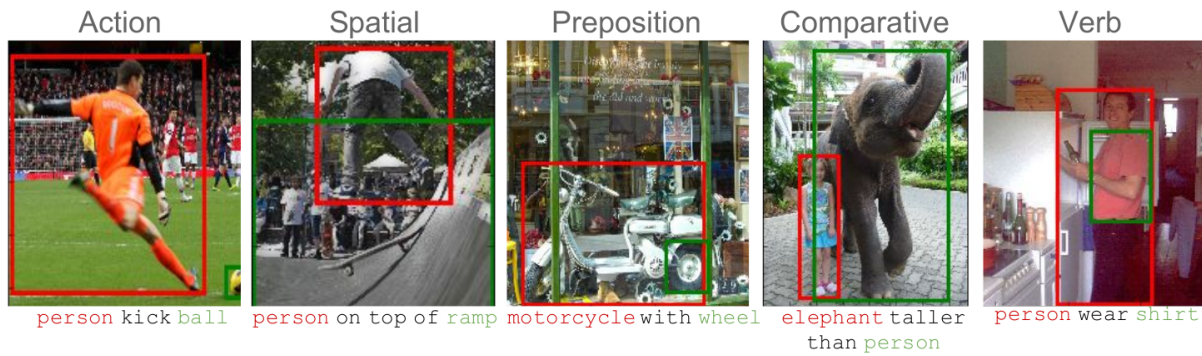
Comme indiqué dans la Section 1.1.3.1, pour ce cas d'usage la configuration spatiale est surtout utile pour l'appariement des parties des objets, à moins de vouloir reconnaître des postures particulières (ou des motifs particuliers sur les vêtements, mais cela nécessiterait une résolution supérieure). D'autres caractéristiques plus pertinentes peuvent alors être utilisées pour la reconnaissance, comme la couleur ou la forme des parties. La méthode de reconnaissance proposée est détaillée dans le Chapitre 6 de ce manuscrit. Par ailleurs, l'entreprise Magellium dispose également de telles données, mais sans annotations de détection ou de segmentation. Notre approche pourra être évaluée sur ces données en la plaçant en sortie d'une étape de détection ou segmentation des objets d'intérêt (personnes par exemple).

---

14. <https://github.com/ishann/AeroScapes>



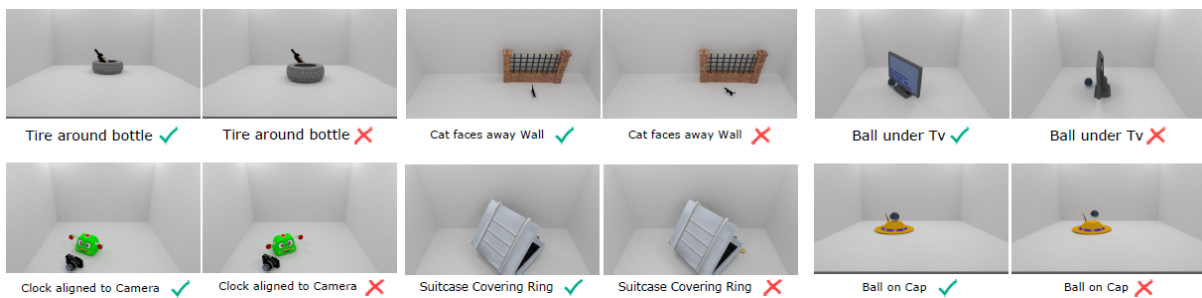
(a) Exemples d'images issues du jeu de données *SUN09* [14], avec annotations issues de [63].



(b) Exemples d'images issues du jeu de données *Visual Relationship Dataset (VRD)* [68].



(c) Exemples d'images issues du jeu de données *SpatialSense* [119].



(d) Exemples d'images issues du jeu de données *Rel3D* [42].

FIGURE 1.7 – Exemples de différents jeux de données contenant des annotations de relations spatiales.



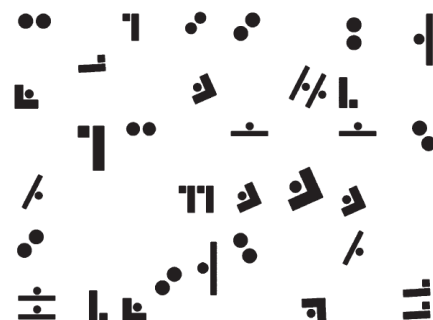


FIGURE 1.8 – Image synthétique utilisée dans [111] pour la reconnaissance de configurations spatiales entre deux objets.

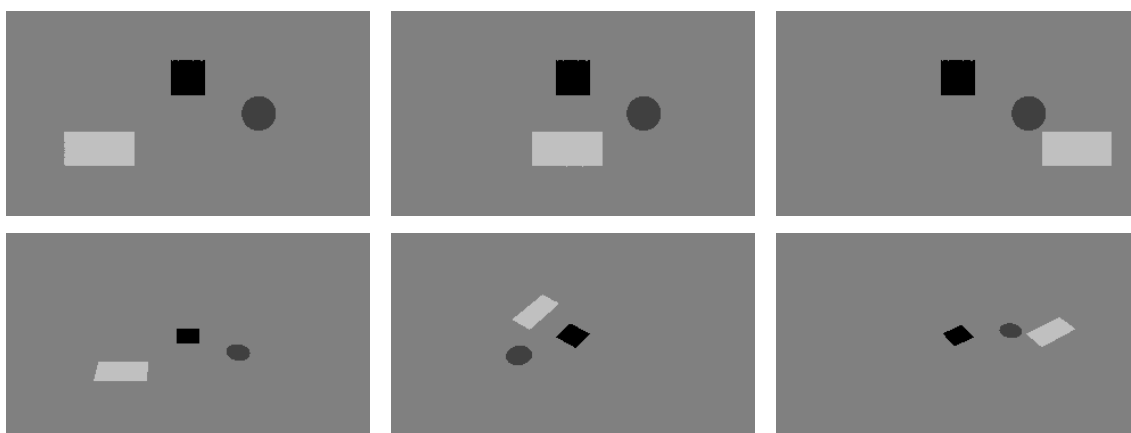


FIGURE 1.9 – Extraits d’une séquence simulée avec *Blender*, avec trois objets dont un mobile, selon deux points de vue 3D différents (vue de dessus et vue oblique avec rotation autour de la scène).

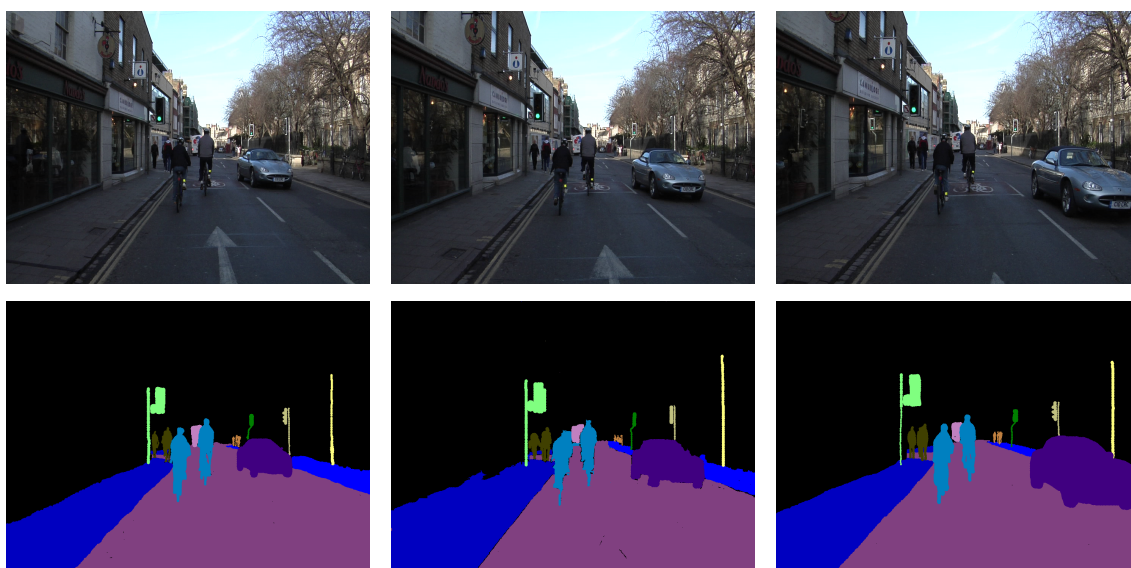


FIGURE 1.10 – Extraits de la séquence *CamSeq01* [37] : images originales et images segmentées, avec sélection de 12 objets d’intérêt.

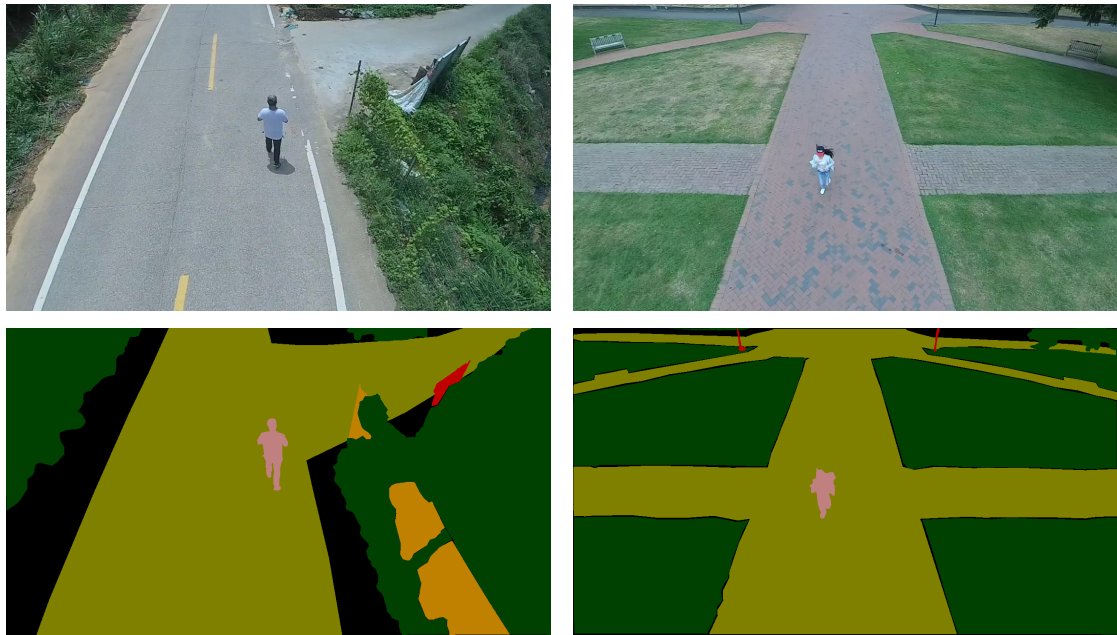


FIGURE 1.11 – Exemples d’images extraites des séquences 000002 et 040000 du jeu de données *AeroScapes* [87] : image originale et image segmentée fournie.



FIGURE 1.12 – Extraits de la séquence 000002 du jeu de données *AeroScapes* [87] : parties de l’image originale obtenues grâce à la segmentation fournie et segmentations basées sur la couleur obtenues par notre méthode, après filtrage des petites composantes.



FIGURE 1.13 – Extraits de la séquence 040000 du jeu de données *AeroScapes* [87] : parties de l’image originale obtenues grâce à la segmentation fournie et segmentations basées sur la couleur obtenues par notre méthode, après filtrage des petites composantes.



## Chapitre 2

# Configuration spatiale entre deux objets

---

2.1	État de l'art . . . . .	36
2.1.1	Descripteurs de relations spatiales . . . . .	36
2.1.1.1	Modèles topo-directionnels élémentaires . . . . .	36
2.1.1.2	Modèles directionnels : descripteurs de position relative . . . . .	38
2.1.1.3	Normalisation et comparaison de descripteurs . . . . .	39
2.1.2	Évaluation de relations spatiales en langage naturel . . . . .	40
2.1.2.1	Approches basées sur des descripteurs dédiés . . . . .	41
2.1.2.2	Approches par détection de triplets (sujet, relation, objet) . . . . .	42
2.2	Description de position relative et reconnaissance de relations spatiales avec le bandeau de forces . . . . .	44
2.2.1	Approche proposée . . . . .	44
2.2.2	Le bandeau de forces . . . . .	44
2.2.2.1	Notions sur l'histogramme de forces . . . . .	45
2.2.2.2	Vers le bandeau de forces . . . . .	46
2.2.2.3	Définition du bandeau de forces ( $d\mathcal{FB}$ ) . . . . .	47
2.2.2.4	Propriétés . . . . .	49
2.2.3	Reconnaissance de relations spatiales avec le bandeau de forces . . . . .	51
2.2.4	Expérimentations et résultats . . . . .	53
2.2.4.1	Données . . . . .	54
2.2.4.2	Protocole expérimental . . . . .	55
2.2.4.3	Résultats et discussion . . . . .	57
2.2.4.4	Analyse de l'emploi de différentes forces . . . . .	60
2.3	Conclusion . . . . .	62

---

La prise en compte de la configuration spatiale des objets est un des défis majeurs de la compréhension de scènes. Plusieurs descripteurs existent déjà pour décrire des configurations de deux objets, comme l'histogramme de forces qui est un exemple typique de descripteur de position relative. En calculant l'interaction entre objets pour une force donnée dans toutes les directions, il donne un bon aperçu de la configuration, et possède des propriétés utiles qui peuvent le rendre invariant au point de vue 2D. Considérant que l'utilisation de forces complémentaires devrait améliorer la description de configurations spatiales complexes, nous proposons d'étendre l'histogramme de forces à un panel de forces afin d'en faire un descripteur plus complet. Cela donne un descripteur 2D que nous avons appelé "*bandeau de forces*", qui peut être utilisé en entrée d'un réseau de neurones convolutif (CNN), bénéficiant de leurs puissantes performances, ou réduit en une représentation spatiale plus compacte pour être utilisé dans un autre système. Pour illustrer sa capacité à décrire des configurations spatiales, nous l'avons utilisé pour résoudre un problème de classification visant à discriminer des relations spatiales simples,

mais avec des configurations de complexités variables. Les résultats expérimentaux obtenus mettent en évidence l'intérêt de cette approche, en particulier pour des configurations spatiales complexes.

## 2.1 État de l'art

Un aperçu historique de la recherche sur la description de configurations spatiales entre deux objets est donné dans la Section 1.1.4.2 du Chapitre 1. Nous détaillons ici le fonctionnement de quelques-unes des principales approches.

Lorsque l'on parle de configurations spatiales entre deux objets, on parle en général de relations spatiales en langage naturel. En informatique, les premiers travaux cherchant à les formaliser et les catégoriser sont ceux de Freeman [39], qui proposa 13 relations spatiales élémentaires pouvant être classées selon trois types : topologiques, directionnelles ou distancielles. À ces relations élémentaires, on peut alors ajouter d'autres relations plus spécifiques impliquant la pose ou la forme des objets, comme détaillé dans la Section 1.1.2.3. Des modèles et des relations décrivant plus précisément la topologie ont été proposées par la suite, et sont détaillés dans la Section 2.1.1.1. Puis, dans le but de décrire complètement la configuration, sont apparus des descripteurs directionnels dit de position relative, dont les principales approches sont détaillées dans la Section 2.1.1.2. Bien qu'avant tout directionnels, ceux-ci prennent également en compte la distance entre les objets et leurs formes, voire la topologie pour les modèles les plus récents. Bien que quantitatifs, ces descripteurs peuvent également être utilisés pour évaluer des relations en langage naturel, selon les approches de la Section 2.1.2.1. Enfin, un dernier type d'approches est apparu plus récemment pour reconnaître des relations spatiales sans passer par des descripteurs dédiés, ce qui est abordé dans la Section 2.1.2.2.

### 2.1.1 Descripteurs de relations spatiales

#### 2.1.1.1 Modèles topo-directionnels élémentaires

Afin de décrire avec précision les différentes configurations spatiales possibles entre deux objets, et en particulier leur topologie, plusieurs modèles ont été proposés pour les images 2D. Utilisant les travaux sur le domaine temporel, notamment les intervalles temporels d'Allen [1], les premiers modèles apparus ont utilisé des projections des objets 2D vers le domaine 1D, selon les deux axes de coordonnées, soit en considérant uniquement le début et la fin de chaque objet selon la projection [43], soit avec une quantification en fonction de l'aire des objets selon la projection [56]. Comme les intervalles d'Allen, ces modèles donnent une décomposition en 13 relations différentes selon les deux directions opposées (cf. Figure 2.1), 8 relations étant suffisantes pour une direction. Il est alors possible d'étendre cette description à davantage de directions en projetant suivant d'autres axes, ce qui permet d'intégrer l'aspect directionnel. Cependant, une telle projection ne permet pas de rendre compte de la diversité et de la complexité des configurations topologiques.

Plusieurs modèles ont alors été proposés pour décrire la topologie dans l'espace 2D. Tout d'abord, le modèle *9-intersections* [36] considère pour chacun des objets sa zone intérieure, son contour et sa zone extérieure, et décrit la configuration par les aires de chacune des 9 intersections possibles, regroupées dans une matrice  $3 \times 3$ . Le modèle *RCC8* [95] qui s'est largement répandu ensuite est une adaptation de la décomposition d'Allen dans l'espace 2D. Il ne prend pas en compte la direction mais uniquement la topologie de la relation, ce qui permet de passer à 8 relations seulement (cf. Figure 2.2). En parallèle, Hernández [46] développe un modèle similaire qui inclut l'aspect directionnel, en combinant une relation topologique et une relation d'orientation, obtenue par une décomposition de l'espace en secteurs angulaires (cf. Figure 2.3). Le modèle *RCC* a ensuite été étendu dans *RCC23* [23] afin de pouvoir décrire davantage de configurations, notamment dans le cas où les objets présentent des concavités, ajoutant pour cela 15 relations d'imbrication.

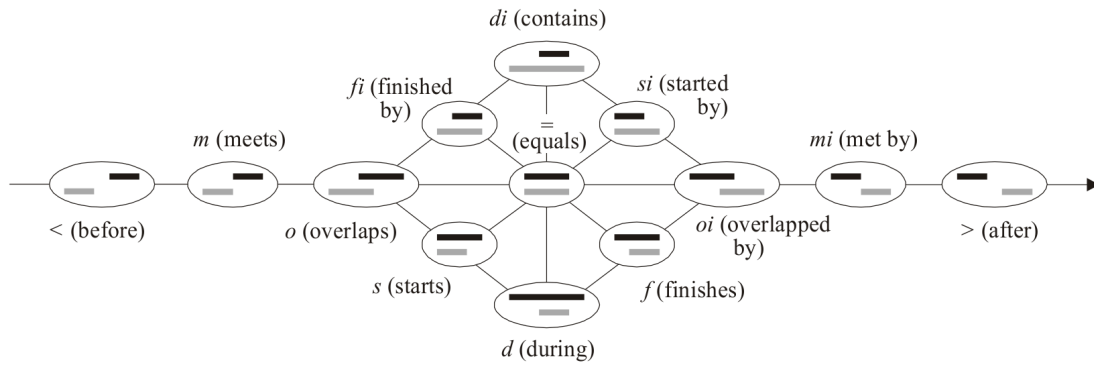
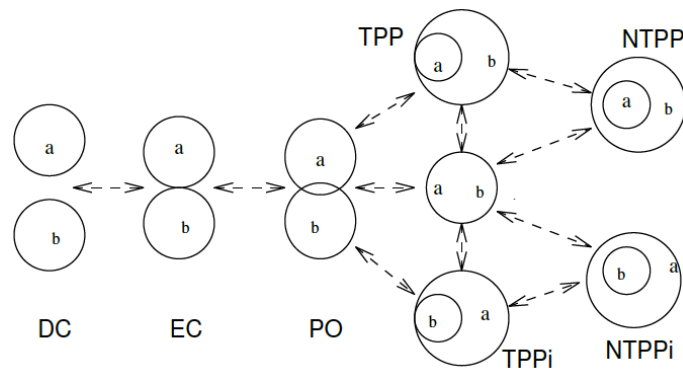
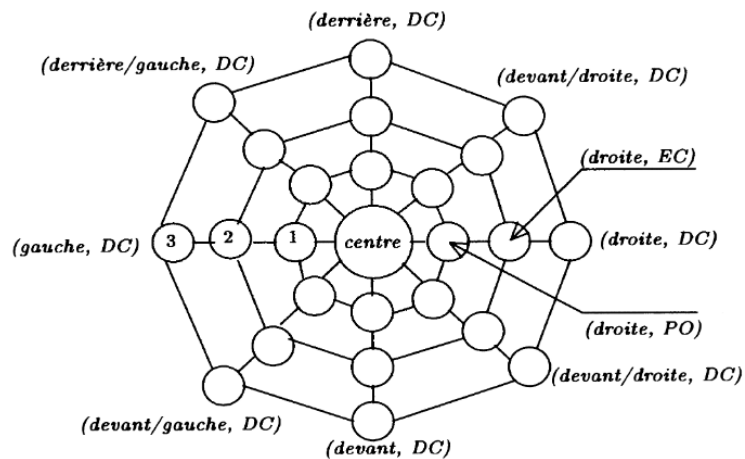


FIGURE 2.1 – Illustration des intervalles temporels d'Allen [1].

FIGURE 2.2 – Illustration des relations spatiales topologiques RCC8 [95]. Les acronymes sont les suivants : DC signifie *disconnected*, EC signifie *externally connected*, PO signifie *partially overlaps*, TPP signifie *tangential proper part* et NTPP signifie *nontangential proper part*.FIGURE 2.3 – Illustration des relations spatiales topologiques (selon 4 possibilités) et directionnelles (en 8 niveaux de granularité) selon Hernández [46], en utilisant les mêmes acronymes que dans la Figure 2.2 (source : [70]). Le mot-clé *centre* correspond à la fusion des 3 relations topologiques EQ, TPP et NTPP.

Tous ces modèles sont utiles lorsqu'il est nécessaire de décrire précisément la topologie d'une configuration donnée, mais ne suffisent pas pour décrire correctement les aspects directionnels et distanciels de celle-ci. De plus, leur nature qualitative ne permet pas de les utiliser facilement pour comparer des



configurations entre elles dans des chaînes de traitement de reconnaissance de formes. D'autres modèles ont alors été proposés pour cela, en décrivant la configuration selon différentes directions.

### 2.1.1.2 Modèles directionnels : descripteurs de position relative

Après s'être attachés à décrire précisément l'aspect topologique d'une configuration spatiale entre deux objets (cf. Section 2.1.1.1), les chercheurs travaillant sur ce sujet se sont intéressés davantage à l'aspect directionnel et à l'aspect distanciel. Hernández est un des premiers à avoir introduit la direction dans un modèle de description de configuration, en combinant une relation topologique et une relation d'orientation, donnée par une quantification de l'angle entre les deux objets [46]. Plutôt que de donner une seule orientation pour décrire la relation directionnelle, les approches suivantes ont alors cherché à évaluer l'ensemble des orientations, de façon graduelle, en calculant diverses mesures suivant cette orientation, obtenant ce qu'on a appelé des descripteurs de position relative. Les premiers proposés se sont focalisés sur cet aspect directionnel en délaissant l'aspect topologique, bien qu'ils donnent une idée de la configuration et de la forme des objets. La prise en compte de la distance permet alors d'obtenir une évaluation encore plus fine et une description plus significative de la configuration. Puis les approches plus récentes ont proposé des solutions pour combiner ces descriptions avec des modèles topologiques, permettant de décrire complètement la configuration.

L'évolution de ces descripteurs appelés descripteurs de position relative est retracée ici, ainsi que celle des paysages flous qui leur est intimement liée. Par ailleurs, une synthèse comparative des descripteurs de position relative est proposée dans [83], tandis que [17] met en avant la dualité entre relations spatiales et position relative. Il est bon de noter aussi que ces approches peuvent généralement être utilisés comme descripteurs de forme (ou "signatures"), en les calculant entre un objet et lui-même.

### Premiers modèles directionnels

Les descripteurs de position relative sont des modèles directionnels dans le sens où ils décrivent la configuration spatiale pour chaque direction de la scène, en calculant diverses mesures suivant cette direction. Ils sont alors représentés sous la forme d'histogrammes représentant ces mesures en fonction de la direction, par secteur angulaire, la représentation polaire étant la plus adaptée. Les approches diffèrent donc principalement par la nature des mesures qui sont effectuées. Ces descripteurs quantitatifs restent assez facilement interprétables et peuvent être exploités pour évaluer des relations spatiales directionnelles en langage naturel, selon différentes méthodes présentées dans la Section 2.1.2.1. Ils peuvent également être utilisés tels quels pour comparer deux configurations, en comparant simplement leurs histogrammes.

Le premier descripteur de position relative a été proposé en 1994 dans [81]. Baptisé "histogramme d'angles", il est calculé en mesurant l'orientation de chaque couple de points entre les deux objets (i.e., de chaque droite formée en prenant un point dans chaque objet), et en comptabilisant le nombre de couples de points par orientation, ou plutôt par secteur angulaire. Cependant, cette approche était assez coûteuse en temps de calcul pour les matériels de l'époque, et est sujette à des discontinuités dues à la discrétisation des objets et des directions.

L'histogramme de forces proposé ensuite est une généralisation de l'histogramme d'angles, qui offre aussi une optimisation intéressante en termes calculatoires. Il a été développé conjointement par P. Matsakis et L. Wendling, qui l'ont d'abord introduit dans leurs thèses de doctorat respectives [71, 117], puis présenté plus largement dans [79]. Sa définition mathématique est donnée dans la Section 2.2.2.1. Il repose sur deux évolutions majeures : le découpage des objets en segments plutôt qu'en points, et la définition d'une force d'interaction entre ces segments, qui est intégrée sur l'ensemble des paires d'une direction donnée (ou d'un secteur angulaire), profitant pour cela des bonnes performances du calcul intégral. En calculant l'interaction entre les objets pour chaque direction, il combine l'aspect directionnel et des mesures de distance, ce qui permet de tenir compte de la forme et de la dimension des

objets. De plus, l'histogramme de forces bénéficie de plusieurs propriétés intéressantes. En particulier, il peut être rendu invariant aux variations de point de vue dans l'image 2D, en utilisant pour cela une étape de normalisation [53]. Il est impliqué dans plusieurs domaines d'application, tels que la mise en correspondance de scènes [9] ou la recherche d'images basée sur le contenu [109, 18]. Par ailleurs, une implémentation de ce descripteur pour des données 3D a été proposée dans [86].

### Modèles directionnels intégrant la topologie

L'histogramme de forces est très pratique mais il ne décrit pas précisément la topologie de la configuration et ne permet donc pas de décrire des configurations complexes. Une solution est alors de combiner une description directionnelle avec une description topologique (cf. Section 2.1.1.1), s'inspirant ainsi du modèle de Hernández [46]. Suivant cette idée, les histogrammes d'Allen [70, 78] sont issus de la combinaison de l'histogramme de forces avec les intervalles d'Allen, en utilisant pour cela 13 histogrammes correspondant aux 13 relations d'Allen, et en catégorisant les segments selon leur relation d'Allen. Cependant, cette première solution ne permet de décrire que des configurations d'objets de formes convexes, et ne permet pas toujours de transcrire la topologie globale de la configuration.

En parallèle, le R-histogramme [115] et le R\*-histogramme [116] ont été proposés. Il s'agit d'extensions de l'histogramme d'angles et de l'histogramme de forces, qui considèrent à la fois la direction et la distance pour produire un histogramme bidimensionnel, et qui exploitent également une information qualitative concernant le chevauchement des objets. Cette solution a été utilisée pour une tâche de recherche d'images, mais plusieurs défauts n'ont pas motivé de développements plus poussés.

Plus récemment, une autre solution plus complexe a été proposée avec le  $\phi$ -descripteur [76], afin de décrire des configurations plus variées comprenant des objets non convexes. Celui-ci se base aussi sur un ensemble d'histogrammes de forces obtenus avec une catégorisation des segments, mais avec une catégorisation plus fine selon 24 catégories rangées en 10 groupes, donnant 10 histogrammes, plutôt que selon les 13 relations d'Allen. Chaque valeur d'un histogramme correspond alors à l'aire d'une région bien précise de la configuration, dont les limites dépendent de la direction observée. Cet ensemble est complété par 3 autres histogrammes de forces dédiés à des fonctions particulières, ainsi que par un ou plusieurs histogrammes dit de longueurs, et enfin les aires des deux objets. Ce descripteur bénéficie de propriétés intéressantes face aux transformations affines [72] et fournit un cadre générique pour évaluer un grand nombre de relations spatiales, en extrayant un ensemble d'opérateurs flous dédiés [75]. Il est également généralisable à des données 3D, où les aires sont alors remplacées par les volumes des régions. Cependant, sa dimensionnalité le rend difficile à exploiter pour décrire une scène comportant un nombre important d'objets.

Enfin, on peut aussi noter une autre approche de description de la position relative avec le *Radial Line Model* [102]. Dans cette solution, la position des objets n'est plus analysée pour un objet par rapport à un autre, mais pour chaque objet par rapport à un point de référence déterminé selon des caractéristiques topologiques entre les deux objets, en utilisant le formalisme 9-intersections proposé par [36]. La description se fait alors en comptabilisant la proportion de pixels de chacun des deux objets selon un échantillonnage de directions, ce qui donne deux histogrammes différents. Cependant, cette description ne décrit pas non plus précisément la topologie de la configuration en comptant sur un unique histogramme. Il serait donc intéressant de la combiner avec une approche intégrant mieux la topologie, comme les histogrammes d'Allen ou le  $\phi$ -descripteur.

#### 2.1.1.3 Normalisation et comparaison de descripteurs

Afin de comparer entre elles des configurations spatiales de deux objets, il est nécessaire de définir les méthodes de comparaison de leurs descripteurs. Par ailleurs, comme introduit dans la Section 1.1.2.4, un point important est de pouvoir prendre en compte ou pas les variations de point de vue de la scène dans cette comparaison. En effet, les descripteurs présentés n'intègrent pas nativement une telle inva-



riance, donc celle-ci ne peut s’obtenir que lors de la comparaison (cf. Section 1.1.4.4). Nous détaillons ici les méthodes de comparaison et de normalisation possibles pour l’histogramme de forces.

La comparaison d’histogrammes est un sujet d’étude à part entière, qui fait l’objet d’une littérature importante et touche de nombreux domaines, avec des formulations variées. Plusieurs synthèses des différentes mesures existantes sont disponibles, par exemple dans [11]. De façon générale, il existe deux méthodes principales pour comparer des histogrammes : la comparaison *bin-to-bin*, comparant chaque intervalle (*bin*) indépendamment des autres, et la comparaison *cross-bin*, comparant deux à deux l’ensemble des intervalles, de façon matricielle. Chacune de ces deux méthodes peut alors s’employer en tenant compte de l’ordre des intervalles ou pas, et en autorisant des décalages et déformations ou pas.

Dans le cas de l’histogramme de forces, on s’intéresse à une comparaison direction par direction (ou plutôt : secteur angulaire par secteur angulaire), donc uniquement *bin-to-bin*, en tenant compte de l’ordre des intervalles puisque les directions sont ordonnées. Par ailleurs, les variations de point de vue se traduisent sur celui-ci par des décalages ou des déformations [73, 85] (cf. Figure 2.10). Par conséquent, ne pas autoriser les décalages et déformations signifie que l’on n’autorise pas les variations de point de vue, ou que l’on souhaite les intégrer dans le résultat ; et inversement, les autoriser signifie que l’on souhaite se rendre invariant à ces variations, ce qui nécessite des méthodes particulières. De plus, il faut rappeler que l’histogramme de forces est périodique, puisqu’il est défini sur un ensemble de directions. Il nécessite donc des mesures adaptées pour prendre en compte cet aspect, comme celle proposée dans [12].

Une autre solution pour obtenir une invariance aux variations de point de vue est d’exploiter les propriétés de l’histogramme de forces pour le normaliser, afin de pouvoir utiliser des mesures *bin-to-bin* classiques. Ces aspects ont notamment été étudiés dans [73], qui propose une méthode de normalisation et évalue 20 mesures différentes pour comparer les histogrammes normalisés (mais pas la mesure précédente). Les auteurs ont alors retenu trois mesures performantes : un indice de Tversky, qui est en fait l’indice de Ružička (défini dans l’équation 3.1), un indice de Pappis et la corrélation croisée normalisée. À noter que indices de Tversky sont des généralisations de l’indice de Jaccard/Tanimoto (plus connu sous le nom de IoU) ou encore de l’indice de Sørensen–Dice, et ont été introduits pour des valeurs binaires. L’indice de Ružička est simplement l’extension de l’indice de Jaccard/Tanimoto aux valeurs réelles, et a pour complémentaire la distance de Soergel, ou distance de Jaccard pondérée. Plus récemment mais dans la continuité de ces travaux, une méthode formelle de normalisation de l’histogramme de forces a été proposée afin d’obtenir des histogrammes invariants aux similitudes [53].

## 2.1.2 Évaluation de relations spatiales en langage naturel

Les descripteurs précédents peuvent être qualitatifs avec les descripteurs topologiques, quantitatifs avec les descripteurs directionnels, ou combiner ces deux aspects avec les descripteurs topodirectionnels. Cependant, dans tous les cas ils ne donnent pas une description aisément compréhensible par un humain, pour qui une description des relations en langage naturel est préférable, de façon à faciliter la compréhension en réduisant le travail d’abstraction [40]. Ainsi, plusieurs tâches de compréhension de scènes ont pour but de décrire ces scènes en langage naturel, que ce soit par des phrases avec le *captioning*, ou pour répondre à des questions avec le VQA ou le raisonnement visuel. Afin de fournir la meilleure description possible, ces tâches nécessitent alors de pouvoir évaluer la validité d’une relation donnée. Deux types d’approches sont possibles pour cela : exploiter les descripteurs précédents, ce qui nécessite une étape de segmentation, ou utiliser des méthodes automatiques traitant directement les images, avec éventuellement une étape de détection ou de segmentation. Nous détaillons ici ces deux approches. À noter que toutes deux supposent d’avoir défini un vocabulaire des relations spatiales et éventuellement de leurs modulations. Il est alors nécessaire que ce vocabulaire corresponde à la perception que l’on peut se faire des relations spatiales, ce qui fait intervenir la psychologie.

### 2.1.2.1 Approches basées sur des descripteurs dédiés

La première approche pour produire une description des relations spatiales en langage naturel consiste à exploiter les descripteurs de configuration spatiale définis précédemment, soit directement pour les descripteurs topologiques, soit par une évaluation quantitative des différentes relations pour les descripteurs de position relative. Les descripteurs topologiques étant plus spécifiques, et déjà qualitatifs par nature, nous nous intéressons ici à la traduction de descripteurs de position relative, qui sont avant tout des descripteurs directionnels quantitatifs, ainsi qu'aux approches similaires de type "pages flous". Par ailleurs, plusieurs descripteurs et méthodes d'évaluation ont également été proposés pour évaluer des relations spatiales spécifiques, par exemple pour les relations "à travers", "le long de", "entre" [6, 67], "entouré par" [97, 112], ou encore "enlacé par" [22, 17], avec une stratégie inspirée de l'histogramme de forces pour cette dernière.

#### Approches basées sur des descripteurs de position relative

Les descripteurs de position relative étant avant tout des descripteurs directionnels, ils sont surtout utiles pour décrire des relations spatiales directionnelles, comme les quatre directions cardinales typiquement. Il est alors possible de moduler ces relations avec des termes adaptés, comme illustré par la Figure 2.4b pour une relation entre deux points, ou de les combiner pour obtenir des relations comme "dans", "autour" ou encore "entre". Plusieurs solutions sont alors possibles pour traduire les descripteurs de position relative en relations spatiales en langage naturel ou pour évaluer ces relations [27], notamment en utilisant la logique floue qui est bien adaptée pour prendre en compte l'aspect graduel [5, 17]. Nous en résumons ici les principales.

Une première méthode proposant une évaluation des relations directionnelles est celle de Kóczy, qui propose une quantification en fonction de l'aire relative des objets projetés suivant les deux axes de coordonnées [56]. D'autres directions pourraient alors être évaluées en utilisant d'autres axes pour obtenir un descripteur directionnel. L'histogramme d'angles et l'histogramme de forces proposés plus tard ont permis d'obtenir une telle description directionnelle, mais sans l'aspect relatif permettant de quantifier la relation pour chaque direction indépendamment. Pour cela il est alors nécessaire de passer par une évaluation en aval. Plusieurs solutions ont été proposées, et sont applicables à ces deux descripteurs.

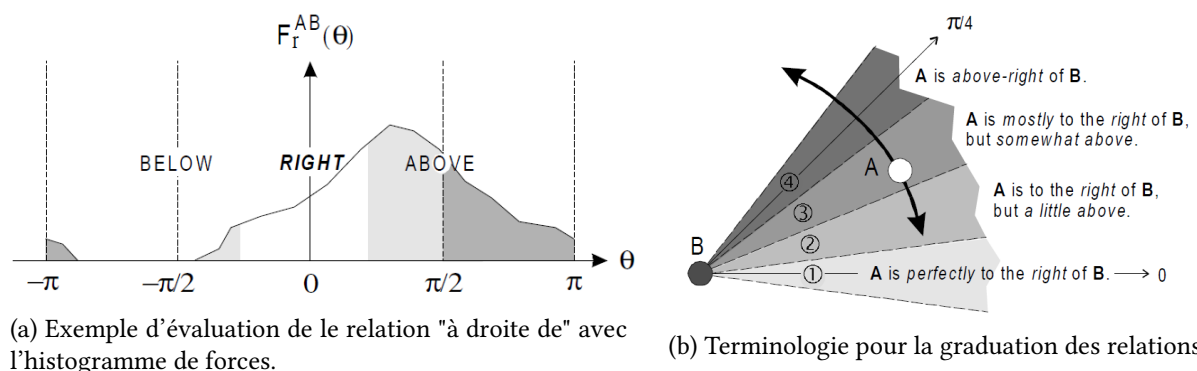


FIGURE 2.4 – Illustrations de la traduction de descripteurs de position relative en relations spatiales en langage naturel (source : [74]).

Deuxièmement, la méthode initiale proposée avec l'histogramme d'angles [81] consiste à comparer le descripteur à un profil flou, tels que ceux présentés dans [17], ce qui est appelé méthode par compatibilité. Elle est similaire à la méthode par agrégation proposée peu de temps auparavant [58], qui consiste à évaluer la validité d'une relation pour chaque couple de points puis à réaliser l'agrégation de ces évaluations par des opérateurs flous. Une autre méthode consiste à utiliser des règles prédéfinies adaptées

aux descripteurs utilisés, grâce à des mesures sur ces descripteurs. [74] en donne un exemple en proposant une méthode de traduction de l'histogramme de forces en relations directionnelles, en considérant l'effet des forces suivant leur direction et la relation considérée, comme illustré sur la Figure 2.4a. Un exemple d'application sur une image aérienne réelle est donné dans la Figure 1.4.

Enfin, il est possible d'utiliser l'apprentissage pour déduire automatiquement la traduction, ce qui a été proposé avec l'histogramme d'angles [55, 114]. Deux méthodes d'apprentissage ont été expérimentées : directement par un réseau de neurones, ou en combinant les sorties de réseaux de neurones dédiés à chaque relation, grâce à des intégrales de Choquet par exemple. Initialement limitée par les performances de calcul, cette méthode peut maintenant être utilisée plus facilement, et permettre de prendre en compte des variations dans les données, à condition de disposer de données d'apprentissage annotées. Ainsi, elle a été utilisée avec l'histogramme de forces, avec un apprentissage par réseau de neurones [8] et nous l'avons utilisée dans nos travaux sur le bandeau de forces, avec un apprentissage par CNN (cf. Section 2.2.3).

### Paysages flous directionnels et *F-templates*

En parallèle des descripteurs de position relative, une autre approche a été développée avec les paysages flous directionnels [3, 4], afin d'évaluer les relations directionnelles, suivant l'idée des *spatial templates* développée auparavant en psychologie [66]. Au lieu de considérer deux objets et de chercher à décrire leur relation, ceux-ci considèrent un seul objet et une relation et définissent la validité de cette relation autour de l'objet (voir Figure 2.5). Pour cela un degré de validité est donné pour chaque pixel de l'image, ce qui en fait une variable floue, d'où la dénomination de paysage flou. Ce modèle basé sur des opérations morphologiques a permis d'obtenir des résultats intéressants dans diverses applications telles que la segmentation cérébrale par IRM [24] ou la détection de bâtiments en imagerie satellitaire [89]. Cependant, cette méthode est relativement coûteuse en temps de calcul. Aussi, une amélioration de celle-ci utilisant l'histogramme de forces a été proposée avec les *F-templates* (ou champs de forces) [77, 80], montrant la dualité entre ces deux concepts.

Par ailleurs, l'apprentissage a également pu être utilisé pour générer des paysages flous "génériques", *i.e.*, communs à tous les objets, grâce à un entraînement sur une base annotée avec des triplets (sujet, relation, objet), pour une tâche de *retrieval* et 11 relations spatiales différentes [69] (cf. Section 2.1.2.2). Enfin, plus récemment, des paysages flous spécifiques ont été proposés pour l'évaluation de la relation spatiale d'enlacement [21]. Néanmoins, toutes ces approches ont toujours le défaut de ne pas bien prendre en compte la forme des objets.



FIGURE 2.5 – Exemples de paysages flous pour la relation "à droite", par rapport à plusieurs objets de référence (source : [4]).

#### 2.1.2.2 Approches par détection de triplets (sujet, relation, objet)

Plutôt que de passer par une représentation des relations spatiales définie "manuellement", avec des descripteurs explicites, une autre approche pour évaluer des relations spatiales consiste à rechercher directement des triplets (sujet, relation, objet). Cette approche permet notamment d'intégrer la sémantique, en considérant des objets bien précis plutôt que des objets quelconques représentés par des masques de segmentation, ce qui permet d'exploiter des a priori sur leur co-occurrence et leurs

relations. Par exemple, un livre aura plutôt tendance à se trouver sur une table plutôt que dessous ou dans un four. L'évaluation peut alors se faire soit à partir des objets segmentés, soit à partir de simples détections (rectangles englobants et étiquettes), soit directement sur les images brutes, par apprentissage typiquement. Ces solutions nécessitent cependant des ensembles d'apprentissage importants, et ne sont pas toujours généralisables à d'autres données suivant les éléments qui sont exploités, contrairement aux approches précédentes. Par ailleurs, cette approche considère souvent d'autres relations en plus des relations spatiales, s'inscrivant dans une tâche plus large de détection de relations visuelles (*visual relationship detection*). Cette tâche est aussi similaire à celle de *captioning* ou de description par des phrases visuelles [99], en se restreignant ici à des phrases structurées par un triplet, ainsi que de la reconnaissance d'interaction humain-objet (HOI). Ainsi, plusieurs jeux de données ont été développés spécifiquement pour ces tâches, comme ceux introduits dans la Section 1.1.3.3.

La méthode la plus courante est de passer par une étape de détection des objets, pour laquelle des modèles performants existent depuis plusieurs années, et d'exploiter leurs rectangles englobants, leurs classes sémantiques ou encore leurs apparences pour en déduire le triplet. On peut par exemple citer [69] qui utilise une détection par modèle par parties (*part-based model*) et un apprentissage de "paysages flous" génériques, s'évaluant sur le jeu de données *SUN09* avec de nouvelles annotations de triplets (cf. Figure 1.7a). Quant à eux, [91, 26] utilisent tous deux une détection par Faster R-CNN, puis apprennent à reconnaître la relation à partir d'une combinaison des coordonnées des rectangles englobants et des caractéristiques d'apparence issues du réseau de détection, avec éventuellement un filtrage selon la classe des objets. Ils s'évaluent notamment sur le jeu de données *VRD* (cf. Figure 1.7b). En introduisant le jeu de données *SpatialSense* (cf. Figure 1.7c), [119] évalue ces deux dernières solutions et plusieurs autres, ainsi que deux méthodes comparatives basées uniquement sur les coordonnées des rectangles englobants et/ou les étiquettes des objets, sans prendre en compte leur apparence. Nous avons utilisé une telle solution comme méthode comparative dans nos travaux, en exploitant uniquement les coordonnées des rectangles englobants (méthode *bbox coords*, cf. Section 2.2.4.2). Plus récemment, [35] propose une approche auto-supervisée avec une composante dédiée à chaque type d'attributs, s'évaluant aussi sur *VRD* et *SpatialSense*. Enfin, la solution de [88] passe par une segmentation et apprend à reconnaître les relations à l'aide d'un réseau de neurones directement sur les masques, dans le but de constituer un index pour une tâche de recherche dans une base. Nous avons également utilisé une telle solution comme méthode comparative dans nos travaux (méthode *bbox image*).

Quelques solutions récentes ont poussé encore plus loin l'automatisation, en cherchant à détecter directement les triplets à partir des images brutes. Pour cela, l'idée est d'apprendre une représentation conjointe des objets et de leur relation, typiquement en entraînant un CNN sur un jeu de données avec de telles annotations de triplets. Cette méthode a également l'avantage de détecter en une seule passe l'ensemble des triplets présents dans l'image, plutôt que de traiter séparément la détection des objets et la description des relations. Cependant, elle nécessite des ensembles d'apprentissage encore plus importants, avec des triplets variés pour pouvoir généraliser correctement, ce qui peut être difficile à obtenir. Ainsi, les premières tentatives [45, 57] sur les jeux de données *SUN09* et *Visual Genome* n'ont pas permis de valider l'applicabilité de l'approche. Mais d'autres jeux de données plus récents et plus importants pourraient permettre cela, comme *SpatialSense* [119] ou *Rel3D* [42] qui sont dédiés aux relations spatiales et ont été construits avec de nombreuses variations intra-classe, notamment de point de vue 3D (cf. Section 1.1.3.3).

Enfin, une autre tâche proche de celle-ci est la reconnaissance d'interaction humain-objet (HOI), qui peut passer par l'exploitation des relations spatiales. Ainsi, [120] propose une approche réalisant conjointement la détection des éléments, de leurs poses, ainsi que des relations spatiales entre les parties du corps. Celles-ci sont modélisées par des fonctions de potentiels spécifiques à différentes classes d'actions et de poses, qui sont obtenues par apprentissage, en utilisant la pose pour obtenir la décomposition en parties du corps. Ce type d'approches peut avoir un intérêt également pour d'autres cas d'usage que sur des personnes, en exploitant la détection de pose par exemple.

## 2.2 Description de position relative et reconnaissance de relations spatiales avec le bandeau de forces

### 2.2.1 Approche proposée

Comme les approches traditionnelles, et contrairement aux approches basées sur la détection de triplets, nous considérons la reconnaissance des relations spatiales indépendamment de la reconnaissance des objets, en supposant que la reconnaissance de la relation spatiale ne nécessite pas de connaître la nature des objets. En effet et intuitivement, la relation spatiale 2D entre deux objets non orientés est généralement indépendante des objets considérés (pour savoir si un objet est à gauche ou à droite d'un autre, on n'a pas besoin de savoir ce qu'ils sont), même si dans certains cas il est possible d'exploiter un tel a priori (par exemple, le ciel est souvent en haut de l'image et le sol en en bas, la relation entre un livre et une table est plus souvent "livre sur table" que "livre sous table"). Ainsi, l'approche proposée considère indépendamment la reconnaissance des objets, qui n'est pas traitée ici, et leurs relations, en utilisant des descripteurs dédiés à cette tâche.

L'histogramme de forces a apporté une avancée dans le domaine de la description de relations spatiales, en permettant de décrire plus efficacement des configurations variées, et a pu être utilisé pour de nombreuses applications [80]. Il dépend d'un paramètre de force, qui modélise différentes perceptions de la relation entre les objets, des forces répulsives aux forces attractives [79]. Dans ce contexte, il peut être utile de combiner plusieurs histogrammes afin d'obtenir une meilleure compréhension de la relation, comme initié dans [74] avec deux forces spécifiques :  $F_0$  et  $F_2$ . Ces deux valeurs sont particulièrement intéressantes car elles modélisent l'aspect directionnel uniquement pour la première, indépendamment de la distance, et l'attraction gravitationnelle pour la seconde, ce qui la rend indépendante de l'échelle de l'image après normalisation [79].

Nous proposons d'aller plus loin dans cette idée en considérant un panel d'histogrammes de forces utilisant divers paramètres de force, ce qui en fait un descripteur 2D plus complet que nous avons appelé "*bandeau de forces*". Ce descripteur peut être utilisé comme entrée d'un CNN 2D, afin de bénéficier des bonnes performances de ces classifieurs, et également de générer des caractéristiques plus compactes qui peuvent être utilisées dans d'autres systèmes pour représenter la configuration spatiale. Ainsi, nous proposons d'utiliser l'apprentissage automatique pour traduire ce descripteur en relations spatiales en langage naturel, ce qui en fait une tâche de classification, s'inspirant de [55, 114] qui utilisent l'histogramme d'angles. Ceci rejoint les problèmes de reconnaissance de triplets (sujet, relation, objet) (cf. Section 2.1.2.2) et d'annotation automatique des relations de scène, qui peuvent être résolus en ajoutant une étape de détection et de segmentation en amont de notre processus.

Nous développons deux pistes principales dans cette section :

- le concept de bandeau de forces, extension de l'histogramme de forces, en expliquant comment il peut être calculé à partir de paires d'objets binaires et en donnant certaines propriétés théoriques dont il hérite ;
- la traduction de descripteurs de position relative en relations spatiales en langage naturel, en utilisant une approche de classification.

Utilisés ensemble, ces deux concepts méthodologiques permettent de donner automatiquement une description des relations spatiales entre objets dans une image. De plus, nous donnons également quelques idées sur la façon d'utiliser des descripteurs de configuration spatiale tels que le bandeau de forces dans une tâche de reconnaissance plus large, et sur la façon de réduire le bandeau de forces en "primitives spatiales" plus compactes dans cette optique.

### 2.2.2 Le bandeau de forces

Dans cette section, nous proposons une représentation qui décrit la configuration de deux objets de manière directionnelle, en calculant les interactions entre ces objets selon la direction et différents types



d'interactions. L'originalité de cette approche, qui la différencie des approches habituelles entièrement basées sur les réseaux de neurones convolutifs (CNN), est de générer des caractéristiques performantes dédiées à l'information spatiale. Cette nouvelle représentation appelée "*bandeau de forces*" généralise le concept d'histogramme de forces [79], puisqu'elle l'étend à un panel de forces, d'attraction et de répulsion. Une telle extension est plus expressive que son homologue, car elle peut utiliser des forces différentes et complémentaires pour représenter la configuration. De plus, de par sa nature planaire, le bandeau de forces peut être utilisé comme entrée d'un CNN 2D classique afin de bénéficier de ses bonnes performances. Il peut alors être utilisé comme caractéristiques spatiales dans un système de classification pour améliorer la compréhension spatiale de la scène, soit tel quel soit en le rendant plus compact si besoin.

Plusieurs questions se posent dans ce contexte :

- Une telle représentation peut-elle être utilisée comme entrée d'une tâche de vision par ordinateur, *e.g.*, une tâche de classification ?
- Cette représentation peut-elle être traduite en descriptions en langage naturel (*e.g.*, l'objet A est à gauche de l'objet B) ?
- Cette représentation est-elle robuste à des variations de point de vue (rotations, translations, changement d'échelle) ?
- Cette représentation est-elle robuste à une segmentation bruitée ou imparfaite ?
- Cette représentation peut-elle être facilement calculée sur tout type d'objets ?

Dans nos expériences, nous combinons les deux premières questions en un problème de classification : considérant deux objets et leur représentation, peut-on utiliser un classifieur pour prédire leur relation spatiale en langage naturel ? Nous évaluons la robustesse à des variations de point de vue en s'évaluant sur un jeu de données synthétique contenant de telles variations. Enfin, concernant les deux dernières questions, nous montrons comment cette représentation peut être calculée sur des images naturelles, en utilisant une étape de segmentation en amont, et nous évaluons celle-ci sur un jeu de données public contenant des annotations de relations spatiales.

### 2.2.2.1 Notions sur l'histogramme de forces

L'histogramme de forces est un descripteur de position relative entre deux objets, qui a été introduit en 1999 dans [79]. Dans le but de caractériser et d'évaluer la configuration spatiale directionnelle d'objets binaires, il repose sur la définition d'une force d'interaction entre objets. Concrètement, l'histogramme de forces est calculé en intégrant une force élémentaire  $\varphi_r(d)$ , fonction de la distance entre points, sur toutes les paires de points entre les deux objets (ou un sous-ensemble selon le pas de discrétisation). Il dépend d'un niveau de force  $r$ , qui génère des forces attractives ( $r > 0$ ), répulsives ( $r < 0$ ) ou constantes ( $r = 0$ ). Dans cette section, nous rappelons brièvement les principales définitions et les principes de ce modèle, qui constitue la base de notre approche.

Le calcul de l'histogramme de forces entre deux objets est défini comme suit. Étant donné deux points situés à une distance  $d$  l'un de l'autre, leur force d'attraction est :

$$\varphi_r(d) = \frac{1}{d^r} \quad (2.1)$$

où  $r$  caractérise le type de force utilisée. Au lieu de travailler directement avec toutes les paires de points entre les deux objets, la force d'attraction entre deux segments unidimensionnels est considérée. Soient  $I$  et  $J$  deux segments d'une droite de direction  $\theta$ ,  $D_{IJ}^\theta$  leur distance et  $|\cdot|$  la longueur d'un segment. La force d'attraction  $f_r$  du segment  $I$  par rapport au segment  $J$  est donnée par :

$$f_r(I, J) = \int_{D_{IJ}^\theta + |J|}^{|I| + D_{IJ}^\theta + |J|} \int_0^{|J|} \varphi_r(u - v) dv du. \quad (2.2)$$

Étant donnés deux objets binaires  $A$  et  $B$ , une droite orientée de direction  $\theta$  dans l'image forme deux ensembles de segments appartenant à chaque objet :  $\mathcal{C}_A = \cup\{I_i\}_{i=1..n}$  et  $\mathcal{C}_B = \cup\{J_j\}_{j=1..m}$ . L'attraction mutuelle entre ces segments est définie comme :

$$F_r^{AB}(\theta, \mathcal{C}_A, \mathcal{C}_B) = \sum_{I \in \mathcal{C}_A} \sum_{J \in \mathcal{C}_B} f_r(I, J). \quad (2.3)$$

Alors, l'ensemble des droites parallèles de direction  $\theta$  parcourant l'ensemble de l'image, noté  $\mathcal{C}_\theta$ , donne l'attraction globale  $F_r^{AB}(\theta)$  entre  $A$  et  $B$  selon la direction  $\theta$ . La Figure 2.6 illustre ce processus pour une direction donnée. Enfin, l'histogramme de forces  $\mathcal{F}_r^{AB}$  est obtenu en calculant  $F_r^{AB}$  sur un ensemble de directions  $\theta \in [0, 2\pi[$ , résumant la position relative d'un objet binaire  $A$  (communément appelé l'argument) par rapport à un objet binaire  $B$  (le référent) de manière circulaire.

Il est également intéressant de noter que si un histogramme de forces est calculé entre un objet et lui-même (*i.e.*, en calculant  $\mathcal{F}_r^{AA}$ ), alors il peut être considéré comme un descripteur de forme de cet objet.

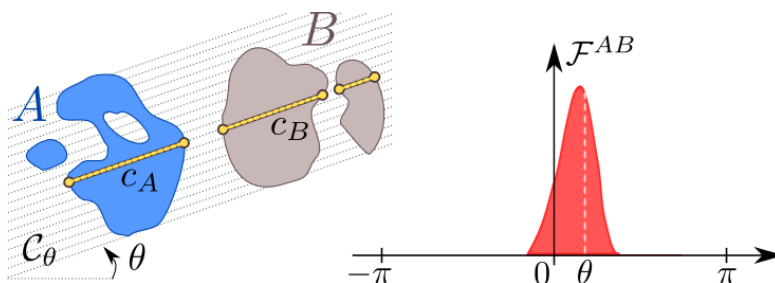


FIGURE 2.6 – Illustration du calcul de l'histogramme de forces : la force d'attraction entre  $A$  et  $B$  selon la direction  $\theta$  est l'intégrale des forces calculées sur les coupes longitudinales ( $\mathcal{C}_A, \mathcal{C}_B$ ) [79].



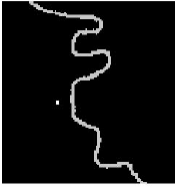
D'un point de vue calculatoire, le calcul d'un histogramme de forces repose sur l'ensemble de toutes les lignes parallèles de direction  $\theta$ , de la même manière que d'autres transformations populaires en traitement d'image (par exemple la transformée de Radon ou la  $\mathcal{R}$ -transform). L'algorithme de l'histogramme de forces s'exécute en  $\Theta(k n \sqrt{n})$ , où  $k$  désigne le nombre de directions discrètes et  $n$  le nombre de pixels [79]. En pratique, le temps de calcul dépend de la complexité des formes considérées (convexité, objets fragmentés, etc.), et dans le pire des cas (rarement atteint pour des objets naturels) une complexité quadratique peut être obtenue. Il a également été démontré que les histogrammes de force peuvent être calculés avec une complexité en  $\Theta(k n \log n)$  lorsque les objets ne se chevauchent pas et qu'une force constante est utilisée [85].

### 2.2.2.2 Vers le bandeau de forces

Deux niveaux de forces sont largement utilisés dans la littérature pour évaluer la relation spatiale entre deux objets avec l'histogramme de forces [79, 74] :

- $r = 0$  repose sur des forces constantes, indépendantes de la distance entre les objets. Dans une certaine mesure, cette approche est basée sur l'utilisation d'un histogramme d'angles isotrope ;
- $r = 2$  repose sur des forces gravitationnelles, où une plus grande importance est donnée aux points les plus proches. Cette approche a la propriété d'être invariante à des changements d'échelle appliqués à toute la scène, après normalisation.

L'évaluation de  $\mathcal{F}_0^{AB}$  donne un aperçu de la configuration spatiale de la scène constituée par  $A$  et  $B$ , mais elle est souvent insuffisante, et peut entraîner une interprétation imprécise dans le cas où la notion de distance doit être prise en compte, comme illustré par les exemples de la Figure 2.7. Un tel comportement peut être corrigé en considérant  $\mathcal{F}_2^{AB}$ , qui se concentre sur des vues rapprochées entre

1	2	3
		

	1		2		3	
	M0	M2	M0	M2	M0	M2
Right	41	<b>64</b>	38	<b>58</b>	03	02
Left	00	00	05	07	51	<b>75</b>
Above	<b>60</b>	40	56	48	37	21
Below	05	05	<b>67</b>	33	<b>59</b>	28

FIGURE 2.7 – Comparaison des scores de confiance obtenus pour les 4 relations directionnelles avec les forces F0 et F2, pour trois scènes différences (source : [79]).

les objets, mais alors une situation complexe peut donner une opinion contradictoire (parfois excessivement pessimiste et parfois excessivement optimiste). En revanche, il a été montré que la combinaison de ces deux types de forces peut fournir un système efficace et robuste pour obtenir une description de la relation spatiale [74].

De plus, les valeurs négatives conduisent à des forces répulsives, ce qui peut être utile pour étudier des formes compactes divisées en plusieurs composantes connexes. Il a été montré que de telles forces intégrées dans un "sac de relations" peuvent apporter un autre point de vue lors d'un processus de classification [20] (avec  $r = -2$ ).

Ainsi, le potentiel de description de  $\mathcal{F}_r^{AB}$  pour une valeur donnée de  $r$  peut être différent selon la complexité des scènes considérées. Dans ce contexte, notre idée est de fournir en une seule représentation, appelée "bandeau de forces" et noté  $\mathcal{FB}^{AB}$  ou  $\mathcal{FB}[AB]$ , une série de  $\mathcal{F}_r^{AB}$  modélisant des forces complémentaires, afin de mieux prendre en compte la complexité d'une situation. Cette représentation peut être utilisée en entrée d'un classifieur (un CNN typiquement, voir Section 2.2.3) pour déduire les relations spatiales entre objets, ou comme primitives spatiales dans un système de reconnaissance, après une étape de compression éventuellement.

### 2.2.2.3 Définition du bandeau de forces ( $d\mathcal{FB}$ )

Soient  $A$  et  $B$  deux objets binaires et soient  $r$  et  $\theta$  deux réels tels que  $r \in [r_s, r_e]$  et  $\theta \in [0, 2\pi[$ . Le bandeau de forces  $\mathcal{FB}^{AB}$  (ou  $\mathcal{FB}[AB]$ ) entre  $A$  et  $B$  est défini par :

$$\mathcal{FB}^{AB} : [0, 2\pi[ \times [r_s, r_e] \rightarrow \mathbb{R}_+$$

$$(\theta, r) \mapsto \mathcal{F}_r^{AB}(\theta) \quad (2.4)$$

En utilisant des données raster, une matrice  $d\mathcal{FB}^{AB}$  est obtenue à partir d'une approximation discrète du bandeau de forces  $\mathcal{FB}^{AB}$ . Considérons  $\Theta = \{\theta_1, \theta_2, \dots, \theta_{|\Theta|}\}$  un ensemble de directions consécutives définies avec un pas constant  $\delta_\theta \in \mathbb{R}$  (i.e.,  $\theta_{i+1} = \theta_i + \delta_\theta$ ,  $\theta_0 = 0$  et  $\theta_{|\Theta|} = 2\pi - \delta_\theta$ ). Et considérons  $R = \{r_s, r_s + \delta_r, \dots, r_e\}$  un ensemble de niveaux de forces entre  $r_s \in \mathbb{R}$  et  $r_e \in \mathbb{R}$ , avec un pas constant  $\delta_r$ . Chaque ligne de la matrice est normalisée par sa propre somme afin d'assurer la même importance pour chaque force. Alors,  $d\mathcal{FB}^{AB}$  est défini comme suit, avec  $r_j = r_s + j\delta_r$  :

$$d\mathcal{FB}^{AB} = \begin{matrix} \left| \begin{array}{cccc} \mu_{\theta_0, r_s} & \cdots & \mu_{\theta_0 + i\delta_\theta, r_s} & \cdots & \mu_{\theta_{2\pi - \delta_\theta}, r_s} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \mu_{\theta_0, r_j} & \cdots & \mu_{\theta_0 + i\delta_\theta, r_j} & \cdots & \mu_{\theta_{2\pi - \delta_\theta}, r_j} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \mu_{\theta_0, r_e} & \cdots & \mu_{\theta_0 + i\delta_\theta, r_e} & \cdots & \mu_{\theta_{2\pi - \delta_\theta}, r_e} \end{array} \right| \end{matrix} \quad (2.5)$$

et

$$\mu_{\theta_i, r_j} = \frac{\mathcal{F}_{r_j}^{AB}(\theta_i)}{\|\mathcal{F}_{r_j}^{AB}(\cdot)\|} \quad (2.6)$$



D'un point de vue calculatoire, le calcul du bandeau de forces discret  $d\mathcal{FB}^{AB}$  peut être opéré naïvement en calculant  $n_r$  fois un histogramme de forces (un par ligne de matrice). Il est cependant bon de noter qu'il est évidemment hautement parallélisable selon  $r$  ou  $\theta$ , en *multithreading* ou dans un environnement distribué. De même, le calcul de l'ensemble des lignes parallèles de direction  $\theta$  peut n'être effectué qu'une seule fois pour l'ensemble des histogrammes.

Le bandeau de forces discret  $d\mathcal{FB}^{AB}$  peut ensuite être codé sous la forme d'une image 2D en échelle de gris, où chaque ligne correspond à une force particulière  $r$ , tandis que chaque colonne représente une direction particulière  $\theta$ . En fixant un petit pas d'échantillonnage pour la direction et le niveau de force, et en raison des propriétés de continuité de  $f$ , il peut être considéré comme une représentation "presque" continue, ce qui en fait une image visuellement lisse. La Figure 2.8 fournit un exemple de bandeau de forces discret représenté sous la forme d'un cylindre 3D, qui est une représentation naturelle en raison de sa périodicité angulaire. La Figure 2.9 donne quelques exemples de bandeaux de forces discrets provenant des différents jeux de données que nous avons utilisés dans nos expérimentations.

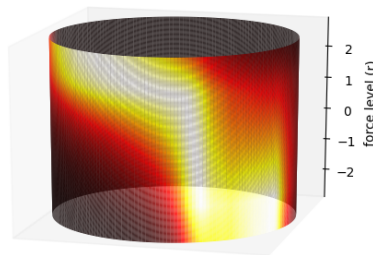


FIGURE 2.8 – Exemple de bandeau de forces discret représenté sous forme de cylindre 3D (correspondant au 2<sup>e</sup> exemple de la Figure 2.9, avec inversion de l'axe vertical). Chaque coordonnée  $z$  renvoie à une force particulière  $r$ , tandis que chaque angle du cercle dans un plan vertical représente une direction  $\theta$ .

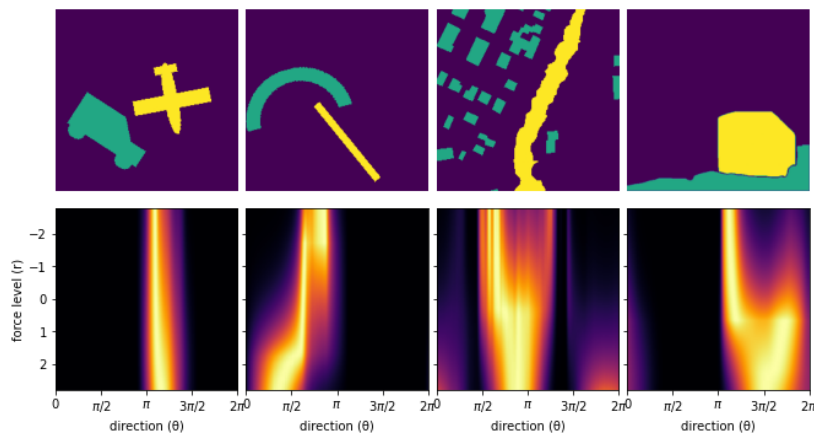


FIGURE 2.9 – Exemples d'images et bandeaux de forces des différents jeux de données considérés. (1<sup>e</sup> ligne) Exemples de formes binaires pour chacun des jeux de données : objets synthétiques (S1), formes géométriques synthétiques (S2), objets segmentés (rivière et maisons) dans une image de télédétection (SIG), objets segmentés (route et bus) dans une image naturelle (S3). Le référent est en jaune tandis que l'argument est en vert. (2<sup>e</sup> ligne) Bandeaux de forces discrets correspondants, décrivant la position relative de l'argument par rapport au référent. Chaque ligne correspond à une force particulière  $r$ , tandis que chaque colonne représente une direction  $\theta$ . Les  $d\mathcal{FB}$  sont représentés sous forme de cartes de chaleur.

### 2.2.2.4 Propriétés

Par héritage de l'histogramme de forces [79], le bandeau de forces a plusieurs propriétés bien utiles souvent attendues dans les processus de reconnaissance de formes. Nous détaillons ici les propriétés les plus importantes. Toutes ces propriétés sont également valables dans le cas discret, avec de faibles variations d'erreur pour la rotation, la réflexion et les transformations affines, dues au pas d'échantillonnage sur les directions.

Par ailleurs, il est important de noter que toutes ces propriétés sont basées sur les cinq propriétés axiomatiques définissant ce que l'on attend d'une relation spatiale [79, 85] : les objets peuvent être assimilés à des points s'ils sont suffisamment éloignés ; les relations directionnelles ne sont pas sensibles aux changements d'échelle ; toutes les directions ont la même importance ; le principe d'inverse sémantique est vérifié (*i.e.*, l'objet A est à gauche de l'objet B autant que B est à droite de A) ; l'évaluation d'une relations spatiale devrait être continue (*i.e.*, un comportement "tout ou rien" n'est pas adapté).

Tout d'abord, il faut rappeler que, de la même manière que l'histogramme de forces, la validité de la définition du bandeau de forces dépend de la valeur de  $r$  et de la topologie de  $A$  et  $B$ , comme indiqué dans la propriété suivante. Une telle propriété est utile notamment pour définir des bandeaux sur des objets qui se chevauchent.

#### Propriété 1 (Ensemble de définition)

*La validité de la définition du bandeau de forces dépend de la valeur de  $r$  et de la topologie de  $A$  et  $B$  :*

- *Si les objets  $A$  et  $B$  sont disjoints et non tangents, alors  $\mathcal{FB}^{AB}$  est défini pour n'importe quelles valeur de  $r_s$  et  $r_e$ .*
- *Si les objets  $A$  et  $B$  sont disjoints et tangents, alors  $\mathcal{FB}^{AB}$  est défini pour  $[r_s, r_e] \subset ]-\infty, 2[$ .*
- *Si les objets  $A$  et  $B$  sont non disjoints, alors  $\mathcal{FB}^{AB}$  est défini pour  $[r_s, r_e] \subset ]-\infty, 1[$ .*

*Dit autrement :*

- *Si  $[r_s, r_e] \subset ]-\infty, 1[$ , alors n'importe quel couple d'objets peut être considéré pour calculer  $\mathcal{FB}^{AB}$ .*
- *Si  $[r_s, r_e] \subset [1, 2[$ , alors des objets disjoints et tangents peuvent être considérés.*
- *Si  $[r_s, r_e] \subset [2, +\infty[$ , alors seuls des objets disjoints et non tangents peuvent être considérés.*

**Justification 1** *L'ensemble de définition du bandeau de forces découle directement de celui de l'histogramme de forces.*

Une des propriétés les plus intéressantes du bandeau de forces est qu'il permet de facilement gérer les similitudes (*i.e.*, les translations, les rotations, les changements d'échelle et les réflexions) lorsqu'elles sont appliquées de la même manière aux deux objets considérés, ce qui correspond à une variation du point de vue de la scène. En effet, il est invariant aux translations, il est juste décalé par une rotation ou inversé par une réflexion, et l'effet d'un changement d'échelle mais aussi de toute transformation affine inversible peut être calculé théoriquement, comme exprimé dans [73, 85] pour l'histogramme de forces et dérivé dans les propriétés suivantes pour le bandeau de forces. Les effets de ces transformations sur l'histogramme de forces sont illustrés sur la Figure 2.10. De plus, il est possible de s'affranchir de l'effet de ces transformations en appliquant une normalisation à l'histogramme ou au bandeau, afin de pouvoir comparer les configurations sans tenir compte de la transformation affectant les objets, à condition qu'elle soit la même sur les deux objets. Plusieurs méthodes ont été proposées pour cela dans [53] pour l'histogramme de forces, et sont transposables au bandeau de forces.

#### Propriété 2 (Impact d'une similitude)

*Le bandeau de forces entre deux images affectées par la même similitude peut être déduit du bandeau entre les images originales, avec une modification dépendant de la transformation :*

- *Translation : Le bandeau de forces est invariant à une translation appliquée sur toute l'image. Ainsi, avec  $t_u$  une translation d'un vecteur  $u$  :*

$$\mathcal{FB}[t_u(AB)] = \mathcal{FB}[AB] \quad (2.7)$$

- *Rotation* : Si on applique à l'image une rotation d'angle  $\alpha$ , alors le bandeau de forces est décalé d'un angle  $-\alpha$  :

$$\mathcal{FB}[\text{rot}_\alpha(AB)](\theta, \cdot) = \mathcal{FB}[AB](\theta - \alpha, \cdot) \quad (2.8)$$

- *Réflexion* : Si on applique à l'image une réflexion selon un axe de direction  $\alpha$ , alors le bandeau de forces est inversé et décalé d'un angle  $2\alpha$  :

$$\mathcal{FB}[\text{ref}_\alpha(AB)](\theta, \cdot) = \mathcal{FB}[AB](2\alpha - \theta, \cdot) \quad (2.9)$$

- *Changement d'échelle* : si on applique à l'image une homothétie  $h$ , alors le bandeau de forces est étiré ou comprimé, selon le facteur d'échelle  $k$  de l'homothétie et selon la valeur de  $r$  :

$$\mathcal{FB}[h_k(AB)](\cdot, r) = k^{3-r} \mathcal{FB}[AB](\cdot, r) \quad (2.10)$$

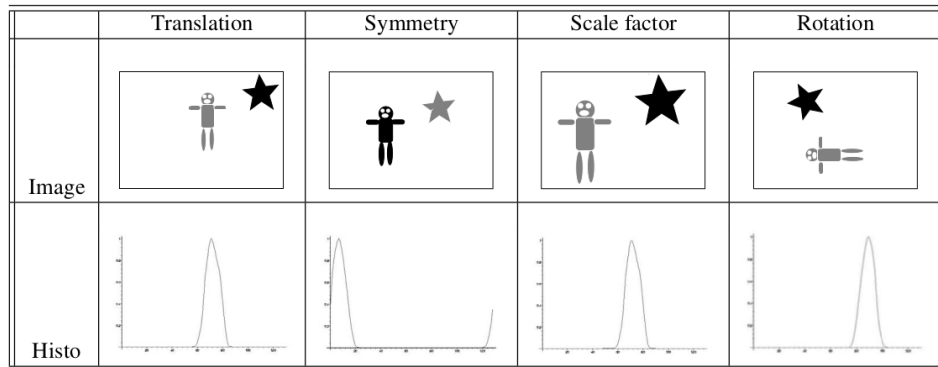


FIGURE 2.10 – Illustration des effets de transformations affines de la scènes sur l'histogramme de forces (source : [110]).

**Justification 2** Un bandeau de forces est défini comme une série d'histogrammes de forces  $F_r$  obtenus pour des valeurs successives de  $r$ . Globalement, une valeur de  $\theta$  a le même impact sur tous les histogrammes définissant le bandeau de forces (le calcul est isotrope) et le calcul de chaque  $F_r$  peut se faire séparément. Les propriétés du bandeau de forces découlent donc naturellement de celles de l'histogramme de forces, qui ont été démontrées dans [79]. En particulier :

- (translation) Le calcul de l'histogramme de forces est indépendant de la position des deux objets dans l'image :  $F_r(\theta, \mathcal{C}_A, \mathcal{C}_B) = F_r(\theta, t_u \mathcal{C}_A, t_u \mathcal{C}_B)$  ;
- (rotation) Le calcul de l'histogramme de forces est isotropique, i.e., pour une direction donnée, il ne dépend pas de la valeur de cette direction en elle-même. Ainsi, si une rotation est appliquée à la scène, il est simplement décalé :  $F_r(\theta, \mathcal{C}_A, \mathcal{C}_B) = F_r(\theta + \theta', \mathcal{C}_{\text{rot}(AB, \theta')})$  ;
- (changement d'échelle) Si une homothétie est appliquée à la scène, alors les forces sur deux segments  $I$  et  $J$  sont multipliées par une valeur qui dépend de  $r$  et du facteur d'échelle  $k$  :  $\forall r \in [r_s, r_e]$  :  $f_r(kI, kJ) = k^{2-r} f_r(I, J)$ , ce qui donne le résultat annoncé après intégration.

### Propriété 3 (Impact de transformations affines inversibles)

Le bandeau de forces entre deux objets affectés par une même transformation affine inversible peut être déduit du bandeau de forces entre les objets originaux, avec une modification dépendant de la transformation :

$$\begin{cases} \mathcal{FB}[\text{aff}(AB)](\theta, r) = |\det(\text{lin})| \cdot |\vec{w}|^{r-1} \mathcal{FB}[AB](\text{angle}(\vec{w}), r) \\ \vec{w} = \text{lin}^{-1} \cdot \vec{\theta} \end{cases} \quad (2.11)$$

où  $\text{aff} = \text{tra} \circ \text{lin}$  est une transformation affine inversible composée d'une translation  $\text{tra}$  et d'une transformation linéaire inversible  $\text{lin}$ , et où  $\vec{\theta}$  est le vecteur unitaire de direction  $\theta$ . Cette formulation inclut les transformations de type similitude de la propriété précédente, ainsi que la transformation d'étirement.

**Justification 3** Cette propriété a été définie et démontrée formellement dans [85] pour l’histogramme de forces. De même que pour la propriété précédente, la justification pour le bandeau de forces découle naturellement de cette démonstration.

Il est également intéressant de noter que le bandeau de forces est comme l’histogramme de forces une relation orientée entre deux objets, avec un référent et un argument, où chacun interagit avec l’autre de manière symétrique. De ce fait, inverser le rôle de chaque objet revient simplement à prendre la direction opposée, ce qui se traduit par un décalage de  $\pi$  sur l’histogramme ou le bandeau. Cette propriété repose sur la notion d’inverse sémantique, en considérant qu’un objet  $A$  est autant à gauche d’un objet  $B$  que l’objet  $B$  est à droite de  $A$ .

**Propriété 4 (Symétrie)**

Les forces exercées sur un objet par un autre sont les mêmes que celles appliquées par l’autre objet dans des directions opposées :

$$\mathcal{FB}[BA](\theta, \cdot) = \mathcal{FB}[AB](\theta + \pi, \cdot) \tag{2.12}$$

Une autre propriété notable de l’histogramme de forces est sa robustesse au bruit et aux petites déformations, au moins pour les valeurs de force positives. En fait, pour de telles valeurs, l’histogramme consiste en un décompte des pixels appartenant à chaque objet le long de sections orientées, donc ajouter ou supprimer seulement quelques pixels d’une direction aura un impact limité sur ce décompte, relativement au nombre de pixels de l’objet dans cette direction. Concernant la robustesse au bruit, plusieurs expériences ont déjà été menées avec différents types de bruits, et ont abouti à des variations mineures sur les résultats finaux pour chaque tâche [20].

Enfin, une dernière chose à noter est que le bandeau de forces, comme l’histogramme de forces, peut être utilisé comme descripteur de forme en le calculant entre un objet et lui-même, le résultat  $\mathcal{FB}^{AA}$  étant appelé signature de l’objet. Celle-ci est symétrique selon les directions, et peut donc être définie sur un intervalle de directions de taille  $\pi$  uniquement.

**Propriété 5 (Symétrie de la signature)**

Les forces exercées entre un objet et lui-même sont les mêmes suivant deux directions opposées :

$$\mathcal{FB}[AA](\theta, \cdot) = \mathcal{FB}[AA](\theta + \pi, \cdot) \tag{2.13}$$

**2.2.3 Reconnaissance de relations spatiales avec le bandeau de forces**

Nous proposons ici un cadre générique pour traduire notre descripteur en relations spatiales exprimées en langage naturel, et en caractéristiques spatiales plus compactes pouvant être utilisées dans un système de classification.

**Traduction en relations spatiales**

Deux options sont possibles pour traduire les descripteurs de position relative en relations spatiales en langage naturel : s’appuyer sur l’apprentissage pour générer automatiquement les transformations comme dans [114], ou utiliser des règles d’évaluation prédéfinies à partir d’analyses théoriques comme dans [74]. Nous nous proposons d’utiliser l’apprentissage pour traduire notre descripteur en relations spatiales en langage naturel, comme cela est fait dans [114] avec l’histogramme d’angles. Avec cette approche, la transformation est apprise à partir d’un jeu de données annoté avec des paires d’objets et leurs relations spatiales.

Étant donné qu’il est planaire, et supposant que la discrétisation est assez fine pour ne pas introduire de discontinuité non désirée, le bandeau de forces est particulièrement adapté aux CNN 2D. Ainsi, il suffit d’utiliser les  $d\mathcal{FB}$  en entrée d’un tel réseau avec les annotations de relations spatiales comme classes pour apprendre la transformation, à condition de disposer de telles annotations.

Habituellement, un tel modèle nécessite beaucoup de données annotées pour généraliser à de nouvelles données, mais les bandeaux ne sont pas aussi variables que les images courantes, *i.e.*, ils n'ont pas une grande entropie, et on ne s'attend pas à des données de test très différentes des données d'entraînement. De plus, les relations spatiales dépendent principalement de la direction et de la distance, donc le  $d\mathcal{FB}$  est particulièrement adapté au problème et ne nécessite pas beaucoup de calcul pour déduire une relation spatiale. Par conséquent, il permet de généraliser sur la classification des relations spatiales tout en apprenant "par cœur" la traduction.

Dans ce contexte, celle-ci peut être apprise avec peu de données d'entraînement et un petit CNN entraîné à partir de zéro, plutôt qu'un CNN de grande taille pré-entraîné sur IMAGENET comme ce qui est souvent fait par défaut, et la fonction de traduction peut être ré-utilisée pour tout  $d\mathcal{FB}$  qui a les mêmes paramètres (le même intervalle de directions et de forces). Ainsi, nos modèles d'expérimentation peuvent être utilisés pour n'importe quelle autre image, en calculant leur  $d\mathcal{FB}$  avec 224 directions et 224 forces comme nous l'avons fait. Ces modèles sont présentés dans la Section 2.2.4.2.

### Extraction de caractéristiques spatiales

En raison de sa redondance et de sa faible entropie, le  $d\mathcal{FB}$  peut également être facilement réduit en une représentation plus compacte grâce à toute méthode de compression ou d'extraction de caractéristiques 2D, comme l'analyse en composantes principales (ACP) ou les auto-encodeurs. L'intérêt est de pouvoir utiliser les informations de cette représentation dans un autre système de classification pour décrire la configuration spatiale, éventuellement en complément d'autres caractéristiques. Là encore, la transformation choisie peut être réutilisée pour tout  $d\mathcal{FB}$  qui a les mêmes paramètres (le même intervalle de directions et de forces), avec une bonne capacité de généralisation.

### Utilisation dans des tâches plus complexes

L'étape suivante après l'extraction de ces primitives serait de les utiliser pour comparer des scènes ou des objets en fonction de leur configuration spatiale, afin de reconnaître une scène spécifique ou de retrouver des scènes similaires dans une base de données par exemple. Cela peut être fait en comparant directement ces primitives si l'on recherche uniquement des configurations et des formes d'objets similaires, mais aussi en les combinant avec d'autres primitives décrivant d'autres caractéristiques telles que la couleur, la texture, etc. De plus, une scène normale est rarement constituée de seulement deux objets, donc la configuration globale ne peut pas être décrite par un seul descripteur dédié aux paires d'objets comme le  $d\mathcal{FB}$ . Cela nécessite une représentation plus large pour inclure plusieurs descripteurs comme celui-ci, comme des sacs de relations ou des graphes. Ces perspectives sont détaillées dans le Chapitre 3 de ce manuscrit.

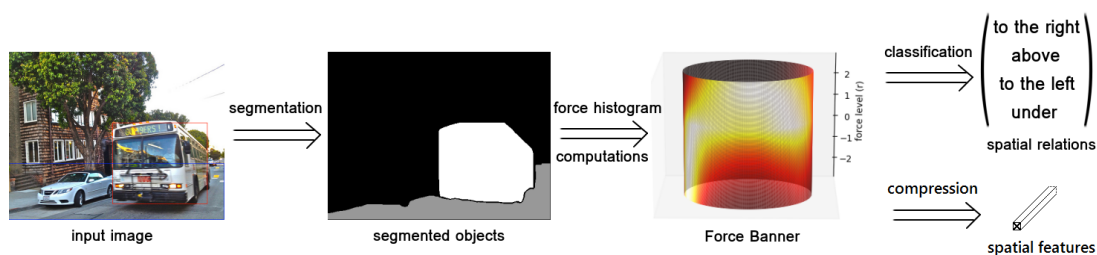


FIGURE 2.11 – Illustration de la chaîne globale appliquée sur le jeu de données *SpatialSense* [119] (S3). De gauche à droite : image originale avec annotations de rectangles englobants, image segmentée avec sélection des deux éléments dans les rectangles englobants, bandeau de force discret (vu comme une carte de chaleur cylindrique à des fins de visualisation), classes de relations spatiales en sortie de la classification, ou vecteur de caractéristiques en sortie de la compression.

### Chaîne globale

L'approche proposée a l'avantage d'être adaptée à tous types d'objets et à toute configuration avec peu de données d'apprentissage, en utilisant le  $d\mathcal{FB}$  comme représentation intermédiaire, ainsi que de généraliser à des associations d'objets et de configurations non vues lors de l'entraînement. Cependant, elle nécessite de travailler avec des objets binaires pour calculer le  $d\mathcal{FB}$ , donc de passer par une étape de segmentation pour utiliser des images naturelles. Dans cette optique, nous proposons une chaîne complète pour extraire des caractéristiques spatiales à partir d'images naturelles, et pour les utiliser pour une tâche de classification (voir Figure 2.11) : (1) segmentation, (2) calcul des  $d\mathcal{FB}$ , (3) traduction en relations/caractéristiques spatiales. Dans nos expérimentations, nous nous sommes concentrés sur la traduction en relations spatiales (voir la Section 2.2.4). Nous avons commencé par considérer cinq relations directionnelles simples, mais cela peut être étendu à davantage de relations si des annotations sont disponibles, en prenant davantage en compte la distance par exemple.

### Padding circulaire

Enfin, un autre point intéressant à noter avec le  $d\mathcal{FB}$  est sa périodicité (selon les directions), ce qui le différencie davantage encore des images naturelles (voir la représentation cylindrique sur la Figure 2.8). Dans les systèmes courants utilisant des CNN sur des images 2D, il est fréquent d'appliquer un remplissage ("*padding*") lors du passage des filtres de convolution sur toute l'image, afin de mieux exploiter le contenu sur les bords. Ceci est réalisé en ajoutant des zéros ou des données interpolées sur les bords de l'image, afin d'obtenir une réponse à ces filtres et ainsi avoir une chance de détecter les motifs recherchés sur les bords également. Or, dans les données périodiques comme le  $d\mathcal{FB}$ , les bords d'une représentation sur une période ne sont pas vraiment les limites des données, puisqu'elles se poursuivent sur l'autre bord en se répétant. Le remplissage est donc plus qu'utile dans ce cas, et devrait être fait en répétant le contenu de l'autre bord, ce que l'on appelle "*padding* circulaire" ou "*looping*". Cette opération est adaptée à tout type de données périodiques, comme les signaux 1D (avec des convolutions 1D) ou les images panoramiques (360°) [105].

### 2.2.4 Expérimentations et résultats

Nous proposons une étude expérimentale basée sur la classification de relations spatiales simples à partir de notre descripteur 2D, de façon similaire à [114]. Pour évaluer le potentiel du bandeau de forces à décrire des configurations spatiales, et particulièrement des relations spatiales complexes, nous considérons la tâche de classification de l'aspect directionnel de la relation spatiale entre deux objets, en choisissant entre les quatre directions principales ("*à droite de*", "*à gauche de*", "*au-dessus*", "*en-dessous*") pour la caractériser, ainsi qu'une cinquième relation "*dans*" dans des expériences supplémentaires.

Pour cela, nous utilisons différents jeux de données d'images contenant des paires d'objets et des annotations de ce type de relations, avec une difficulté variable pour évaluer la relation, et nous calculons leurs représentations  $d\mathcal{FB}$  pour les utiliser en classification. Ci-après nous détaillons le protocole expérimental considéré, les données et les modèles que nous avons utilisés dans notre étude. Puis nous rapportons et analysons nos résultats, et en particulier l'intérêt d'avoir plusieurs niveaux de forces, grâce à une stratégie d'attention visuelle intégrée dans la chaîne.

En fait, cette expérimentation n'est qu'une vérification que le bandeau de forces est capable de distinguer les directions principales, ce qui semble assez évident par définition de l'histogramme de forces. Du point de vue de l'apprentissage automatique, ce descripteur devrait améliorer la généralisation pour cette tâche, en la rendant encore plus facile. Une autre expérimentation possible serait de voir si un tel descripteur peut apprendre à reconnaître des configurations similaires, c'est-à-dire de voir s'il leur donne des représentations similaires.



### 2.2.4.1 Données

Dans cette étude expérimentale, deux types de données ont été utilisées : des données synthétiques et des données réelles. La Figure 2.9 donne des exemples de chacun des jeux de données utilisés, tandis que la Section 1.2.1 détaille le processus de conception de ceux-ci.

#### Images synthétiques

Dans un premier temps, un jeu de données d'images synthétiques a été généré et annoté spécialement pour cette étude. Ce jeu de données est détaillé dans la Section 1.2.1.1. Nommé *2SimpleShapes*, il a été initié avec une première partie constituée d'objets barycentriques, puis agrandi avec des configurations plus variées, ce qui a donné les deux sous-ensembles suivants :

- *2SimpleShapes1* (S1) : 1 000 images de paires d'objets (10 images par paire) de taille  $224 \times 224$ , obtenues à partir de 10 formes différentes représentant des objets basiques (maisons, avions, voitures, etc.) construits à partir de formes géométriques simples (triangles, rectangles, ellipses). Toutes ces paires ont en général des configurations simples puisque tous les objets sont totalement remplis, barycentriques (*i.e.*, construits autour de leur barycentre) et ont des tailles comparables ;
- *2SimpleShapes2* (S2) : 1 280 images de paires d'objets (20 images par paire) de taille  $224 \times 224$ , obtenues à partir de 8 formes géométriques simples (triangles, rectangles, semi-ellipses), petites ou allongées, avec des paramètres variables. Ces paires ont des configurations plus complexes, étant donné qu'elles contiennent des objets de tailles variables et des semi-ellipses, qui ne sont pas barycentriques.

Ce jeu de données a permis une première évaluation quantitative étendue, avec des configurations variées mais limitées en complexité par la taille des images et la forme des objets. Il pourrait ainsi être étendu avec des images plus grandes et des objets plus variés.

#### Images naturelles

Nous évaluons également les performances de notre approche sur des images naturelles pour vérifier comment elle peut être intégrée dans un problème plus concret. Cependant, à notre connaissance il n'existe aucun jeu de données dans la littérature avec à la fois des annotations de segmentation et de relations spatiales. Deux options étaient donc possibles : annoter avec des relations spatiales un jeu de données contenant des objets segmentés, comme nous l'avons fait pour les images synthétiques, ou segmenter un jeu de données contenant des annotations de relations spatiales. Nous avons utilisé les deux approches : pour la première, nous avons découpé en tuiles à différentes résolutions une image de télédétection déjà segmentée, et utilisé les différentes couches pour extraire des paires d'objets, que nous avons annotées comme précédemment ; et pour la seconde, nous avons utilisé le jeu de données *SpatialSense* [119], qui est un jeu de données d'images en couleur contenant des annotations d'objets localisés par leurs rectangles englobants et de relations spatiales entre paires d'objets, que nous avons segmenté selon la méthode présentée dans le Chapitre 4.

Nous avons alors obtenu les deux jeux de données suivants :

- *image SIG* (SIG) : 211 images (190 après rejet des cas ambigus –  $N4$ ) de taille  $224 \times 224$ , contenant des paires d'objets ou d'ensembles d'objets urbains d'une image de télédétection (maisons, routes, rivière, champs...). Ce jeu de données est détaillé dans la Section 1.2.1.2. Il est intéressant de noter qu'il contient des formes composées de plusieurs parties qui ne sont pas connexes, ce qui peut entraîner des situations plus complexes (cf. Figure 2.9, 3<sup>e</sup> colonne). En raison de sa petite taille, cet ensemble de données n'est utilisé que comme cas de test réaliste dans nos expériences ;
- *SpatialSense* [119] (S3) : 2 290 images pour les quatre relations précédentes, et 679 images pour la relation "dans". Ce jeu de données est détaillé dans la Section 1.2.1.3. Ces images sont de taille variable et contiennent de multiples objets de la vie quotidienne, des animaux, des personnes,

mais aussi des "objets" d'arrière-plan (ciel, sol, murs, etc.) Il est à noter que les relations de *SpatialSense* sont conçues pour des modèles prenant en compte les 3 dimensions de la scène, en les donnant du point de vue de l'objet référent (cf. Section 1.1.2.4). Par exemple, un homme vu de face conduira à des relations inversées par rapport aux annotations 2D. De ce fait, ce jeu de données est beaucoup plus difficile à classifier et nécessiterait un traitement particulier, avec une étape d'estimation de pose typiquement.

### 2.2.4.2 Protocole expérimental

#### Modèle retenu

Comme introduit dans la Section 2.2.3, nous suggérons d'utiliser un CNN 2D pour manipuler les bandeaux de forces, étant donné que cette représentation est planaire. Comme précisé également, il n'est pas nécessaire d'utiliser un gros réseau pour l'apprentissage de la transformation des bandeaux en composantes plus simples (relations spatiales en langage naturel ou "primitives spatiales"), étant donné qu'ils présentent déjà clairement l'information utile pour cette tâche.

Nous avons choisi d'utiliser un petit réseau appelé *SmallCNN* et composé de deux couches de convolution, avec activation ReLU, et d'une couche complètement connectée en sortie, qu'on a entraîné à partir de zéro sur les données d'entraînement. La taille du noyau de convolution et le pas lors du parcours de l'image ("*stride*") ont été choisis assez grands pour la première couche (respectivement  $28 \times 28$  et 7 pixels), de façon à capturer suffisamment de contenu dans les bandeaux pour détecter des variations intéressantes. La seconde couche quant à elle a une taille de noyau et un pas plus petits (respectivement  $5 \times 5$  et 2 pixels). 48 canaux sont utilisés dans chacune des deux couches, ce qui fait un total de 127 780 paramètres en comptant la couche finale complètement connectée, pour des images d'entrée de taille  $224 \times 224$  et 4 classes en sortie. De plus, une stratégie de "*padding* circulaire" est utilisée pour les bandeaux de forces, ce qui donne alors 132 772 paramètres (entrées de taille  $224 \times 245$ ).

Nous avons tout de même utilisé un réseau plus grand et pré-entraîné sur IMAGENET à des fins de comparaison, avec le réseau *SqueezeNet*. L'architecture originale de ce modèle est donnée dans la Figure 2.12. Bien que celui-ci ait été entraîné pour une tâche très différente, sur des images naturelles et non des histogrammes, il a été observé qu'un tel pré-entraînement était performant sur des données d'un autre type, par exemple des données audio de type spectrogramme [47]. Ainsi, il a été adapté à la tâche en remplaçant juste la dernière couche de classification en 1 000 classes par une classification en 4 classes (ou 5 pour le test sur S3 avec la relation "dans"), grâce à une couche linéaire complètement connectée. Le réseau ainsi obtenu contient un total de 1 239 500 paramètres, sur des images d'entrée de taille  $224 \times 224$  pour lequel il est optimisé.

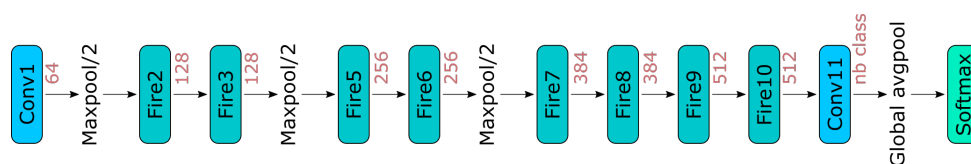


FIGURE 2.12 – Architecture du réseau *SqueezeNet* [51].

#### Méthodes comparatives

Afin de mesurer l'intérêt de notre approche, appelée *dF-banner* ou *dFB*, nous la comparons à plusieurs références :

- *bbbox coords* : classification supervisée basée sur les coordonnées des rectangles englobants des objets, ce qui est similaire à la méthode *2D-only* utilisée dans [119]. Un MLP avec activation ReLU est utilisé pour cela.



- *bbox image* : classification supervisée basée sur l'image binaire, en entraînant un CNN similaire à celui utilisé pour le *dFB*. Comme pré-traitement intuitif pour aider la classification, l'image est réduite au rectangle englobant contenant les deux objets et mise à l'échelle  $224 \times 224$ .
- *F0+F2* : classification supervisée basée sur la combinaison des forces *F0* et *F2*, en dupliquant les histogrammes correspondants et en entraînant un CNN similaire à celui utilisé pour le *dFB*.
- *quadrants* : classification non supervisée basée sur l'image binaire, en divisant l'image en quadrants à partir du barycentre de l'objet référent, et en observant dans quel quadrant l'objet argument est présent. Nous utilisons deux variantes de cette solution : (1) *center* : on calcule le barycentre de l'objet et on prend la décision en fonction de celui-ci; (2) *mask* : on prend une décision pour chaque pixel de l'objet argument et on retient le quadrant avec le plus de votes.

Comme classifieur pour la méthode *bbox coords*, un MLP avec activation ReLU est utilisé comme dans [119] mais avec quelques variations. Dans cette approche, les coordonnées du rectangle englobant de chaque objet sont codées dans des couches linéaires par des vecteurs de taille 512, et fusionnées en un seul vecteur par addition terme à terme, qui est ensuite classifié par un réseau à 2 couches avec 256 unités cachées. Dans notre solution, le réseau prend directement les 8 coordonnées sans les considérer séparément dans la première étape, ce qui permet plus d'expressivité, et il est composé de 4 couches avec 96, 192 et 96 unités cachées, ce qui donne 38 501 paramètres avec 4 classes de sortie.

Pour terminer, dans la méthode *dFB+image* nous utilisons la fusion des sorties des deux modèles entraînés sur les images binaires (*bbox image*) et leurs bandeaux (*dF-banner*), afin d'évaluer s'ils contiennent des informations supplémentaires et complémentaires. Cette étape de fusion est constituée d'une régression linéaire à partir des caractéristiques de la dernière couche avant la décision issue de chacune des deux méthodes. Par ailleurs, une autre option pour le bandeau de forces pourrait être d'utiliser une Analyse en Composantes Principales (ACP) afin de réduire la dimensionnalité de celui-ci, associée à un classifieur plus simple comme un MLP ou un SVM.

### Calcul des bandeaux de forces

Pour être compatible avec les exigences courantes des CNN sans avoir à redimensionner les images, nous avons considéré 224 directions ( $|\Theta| = 224$  et  $\delta_\theta = 2\pi/224$ ) et 224 forces de  $r_s = -2, 8$  à  $r_e = 2, 775$ , avec un pas de 0,025, ce qui inclut les forces *F0* et *F2*. Avec ces paramètres, nous calculons un *dFB* sur des images de taille ( $224 \times 224$ ) en 1.73s en moyenne dans nos expérimentations, avec 8 noyaux CPU (Intel Core i7-8665U 1.90GHz) utilisés en parallèle. La Figure 2.9 présente des exemples des bandeaux ainsi obtenus pour les différents jeux de données. Sur la Figure 2.13, les images et leurs bandeaux de forces sont rangés selon la relation spatiale et le niveau de complexité de cette relation.

### Entraînement des modèles et test

Chacun des modèles est entraîné sur deux jeux de données : le premier est la fusion de *2SimpleShapes1* et *2SimpleShapes2* (S1+S2), et le second est celui issu de *SpatialSense* (S3). En effet, comme indiqué dans la Section 2.2.4.1, le point de vue pris pour l'annotation n'est pas le même pour ces deux jeux de données, donc ils nécessitent des modèles distincts, c'est-à-dire des apprentissages distincts. De la même façon, deux entraînements différents sont réalisés sur S3 : l'un avec et l'autre sans la relation supplémentaire "dans". Sur S1 et S2, l'entraînement est réalisé en excluant les cas totalement ambigus (N4), c'est-à-dire sur les niveaux de difficulté N1+N2+N3, ce qui fait un total de 1993 images. Afin d'éviter un biais d'apprentissage dans l'évaluation tout en utilisant l'ensemble des données pour l'entraînement et le test, une stratégie de validation croisée est utilisée en divisant les données en quatre sous-ensembles qui sont alternativement utilisés comme ensemble de test (25% des échantillons), pour quatre modèles différents entraînés sur la fusion des autres sous-ensembles (les 75% restants). Les mêmes proportions de chaque classe et de chaque niveau de difficulté (pour S1 et S2), sont conservées dans chaque sous-ensemble afin d'avoir une performance comparable. Par ailleurs, afin d'évaluer l'impact du niveau de

difficulté de S1 et S2, les scores ont aussi été calculés sur des sous-ensembles de test correspondant à différents niveaux de difficulté, en utilisant le même apprentissage.

Tous les modèles sont entraînés avec une fonction d'entropie croisée comme fonction de coût, un optimiseur *Adam*, un pas d'apprentissage initial de  $10^{-4}$ , diminué de  $10^{-6}$  par *epoch*, 8 éléments par *batch*, et les valeurs par défaut pour les autres paramètres ( $\beta_1 = 0,9$ ,  $\beta_2 = 0,999$  et  $\epsilon = 10^{-8}$ ). Quant au jeu de données issu de l'image SIG, il est uniquement utilisé comme ensemble de test, ce qui permet d'évaluer la capacité du modèle à généraliser.

### 2.2.4.3 Résultats et discussion

#### Résultats préliminaires

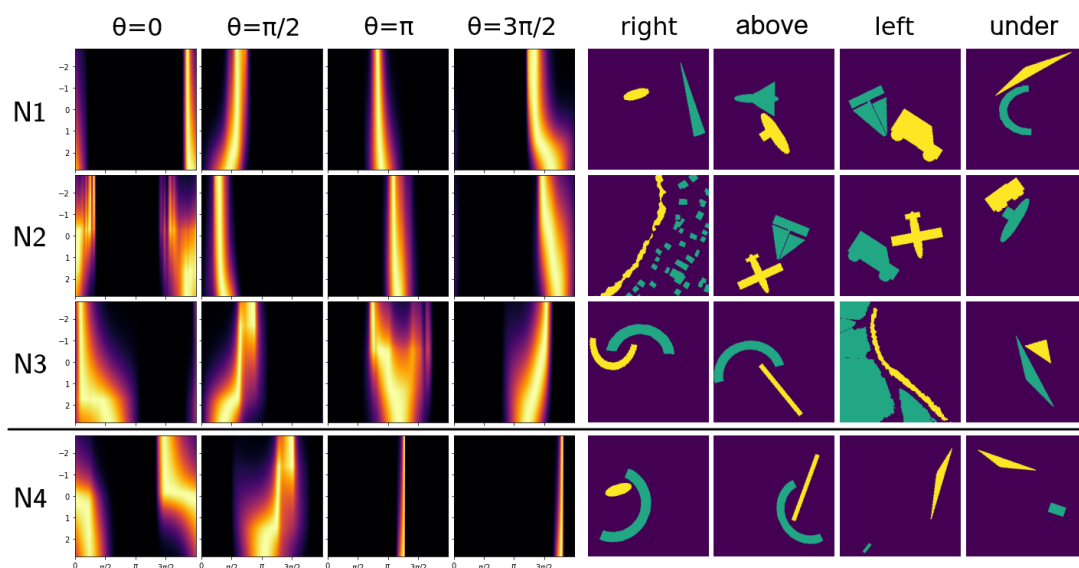


FIGURE 2.13 – Exemples de bandeaux de forces pour les différentes classes (colonnes) et les différents niveaux d'ambiguïté (lignes), de  $N1$  (facile) à  $N4$  (non décidable). Pour chaque classe une direction privilégiée peut être associée ( $0$ ,  $\pi/2$ ,  $\pi$  or  $3\pi/2$ ), ce qui est clair pour les cas simples et devient flou ou trop éloigné de toute direction privilégiée à mesure que la difficulté augmente.

Des exemples illustratifs de bandeaux de force obtenus pour les différentes relations spatiales et les différents niveaux de difficulté sont donnés sur la Figure 2.13. On observe que la direction de la relation est clairement traduite en une direction principale dans le bandeau de forces pour les cas simples (une position sur l'axe horizontal de la représentation 2D), alors qu'elle s'étale sur plusieurs directions pour les cas difficiles, en fonction du niveau de force (coordonnée verticale). Ainsi, il semble facile de prédire la meilleure direction à partir du bandeau dans les cas simples (et aussi dans les cas ambigus où l'intervalle de directions est petit mais ne correspond pas parfaitement à l'une des quatre considérées), où un niveau de force serait en fait suffisant, mais pas dans les cas avec un grand intervalle de directions, où le fait de disposer d'un panel de forces devrait aider à la décision. Dans ce cas, un traitement plus avancé est nécessaire, comme celui que nous proposons avec la classification par CNN.

#### Étude comparative

Les résultats de classification sont rapportés dans le Tableau 2.1 pour le modèle *SmallCNN*, et dans le Tableau 2.2 pour le modèle *SqueezeNet*. Pour chaque méthode et sous-ensemble, la meilleure précision de test sur 50 *epochs* d'entraînement est rapportée, en précisant sa moyenne et son écart-type sur les quatre modèles entraînés avec les quatre partitions de validation croisée. Le taux de bonne classification

(*accuracy*) peut être utilisé comme métrique appropriée ici même si les classes ne sont pas parfaitement équilibrées, puisqu'elles jouent toutes le même rôle.

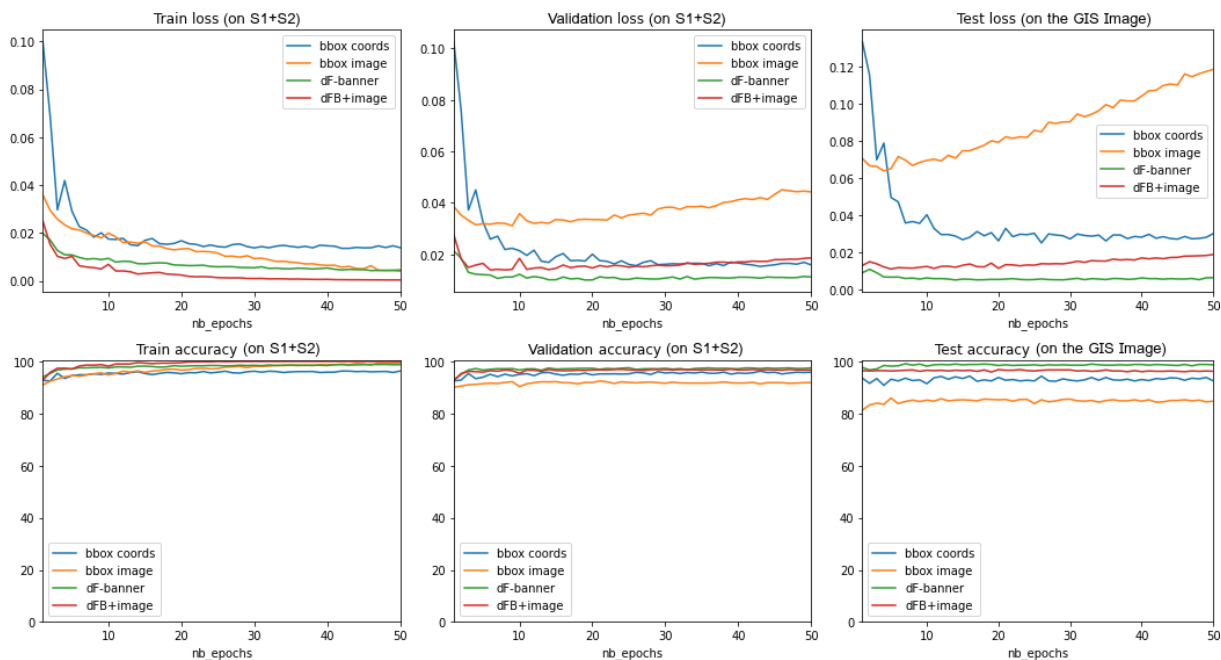
Datasets	<i>F0+F2</i>		<i>dF-banner</i>		<i>bbox image</i>		<i>dFB+image</i>		<i>bbox coords</i>		<i>quadrants</i>	
	OA	STD	OA	STD	OA	STD	OA	STD	centers	masks	OA	STD
<b>Entraînement sur S1+S2</b>												
Test sur S1+S2	97.79	0.82	97.89	0.53	93.08	0.87	97.24	1.15	96.54	0.19	95.48	95.08
Test sur S1	98.96	0.55	98.96	0.45	96.33	1.43	98.61	0.55	97.92	0.30	95.84	95.96
Test sur S2	97.42	0.99	97.32	1.01	91.72	1.79	96.96	1.64	95.83	0.58	95.21	94.41
Test sur S1+S2_N2-3	98.77	0.82	97.75	0.41	95.49	1.71	98.57	1.03	93.03	0.82	87.41	86.54
Test sur S1+S2_N3	97.40	2.00	95.83	0.00	91.67	2.95	96.88	2.69	86.98	1.04	81.41	82.91
Test sur GIS	99.47	0.00	99.47	0.00	87.37	1.36	97.63	0.52	96.19	0.26	96.84	94.74
<b>Test sur S3 (4 classes)</b>									65.68	65.82		
Entraînement sur S3	72.77	2.96	71.35	2.76	66.76	1.97	72.20	2.30	71.78	4.69		
Entraînement sur S1+S2	66.47	1.00	66.19	0.64	64.89	0.82	66.79	0.73	66.70	1.16		
<b>Test sur S3 (5 classes)</b>											non évalué	
Entraînement sur S3	66.81	2.64	65.44	1.91	60.93	1.48	66.74	2.02	64.38	1.42		

TABLEAU 2.1 – Résultats de classification sur les ensembles de test pour les différentes méthodes (taux de bonne classification maximal sur 50 *epochs* – OA en %, et écart-type sur 4 partitions de validation croisée – STD en %), avec le modèle *SmallCNN* pour les données images (4 premières méthodes). Les jeux de données S1 et S2 renvoient à *2SimpleShapes*, SIG à l'image de télédétection et S3 à *SpatialSense*, qui contient la classe supplémentaire "dans". Pour chaque test, le meilleur score est en rouge et gras, les scores proches en rouge.

Tout d'abord, on peut affirmer que la tâche de classification sur des annotations 2D (sur S1, S2 et l'image SIG) est assez simple étant donné qu'elle atteint une très bonne précision de test (plus de 97% avec la meilleure méthode) et qu'elle ne nécessite pas beaucoup d'*epochs* d'apprentissage pour atteindre ces valeurs, comme on peut l'observer sur la Figure 2.14 qui donne un exemple des courbes d'apprentissage. Nous avons réalisé tous les entraînements pendant 50 *epochs* pour tirer le meilleur parti de chaque modèle (en retenant le meilleur score), mais 10 *epochs* semblent suffisantes : au-delà cela n'ajoute que quelques 0,1% aux scores de validation (prédiction sur la partie de test du même jeu de données) et produit même un sur-entraînement pour certaines méthodes, où la précision sur l'ensemble d'apprentissage ne cesse de croître jusqu'à 100% et l'erreur d'entraînement de diminuer, tandis que l'erreur de validation augmente et que la précision de validation reste stable.

Les meilleurs scores sont obtenus par les méthodes *F0+F2*, *dF-banner* et *dFB+image*, c'est-à-dire les méthodes basées sur des combinaisons d'histogrammes de forces et éventuellement de l'image, tandis que la méthode *bbox image* seule est la plus basse, derrière les méthodes *bbox coords* et *quadrants* qui donnent des résultats honorables et sont donc deux méthodes de comparaison valables. Les bons résultats obtenus pour le bandeau de forces associé à un CNN montrent qu'il s'agit d'une solution efficace pour distinguer les relations directionnelles, de même que la combinaison des forces *F0* et *F2*. De plus, la différence entre les scores d'apprentissage (non rapportés ici), de validation (sur S1 et S2) et de test (sur l'image SIG) est vraiment faible pour le *dFB*, ce qui confirme qu'il est capable de généraliser et bien adapté pour simplifier le problème, en extrayant le contenu utile et discriminant de l'image. En particulier, le bon score pour le test sur le jeu SIG montre qu'il est capable de généraliser à des objets constitués de nombreuses sous-parties, contrairement à *bbox image* qui nécessiterait un apprentissage spécifique pour ces images différentes.

Des tests additionnels ont été menés sur des sous-ensembles de S1+S2 afin d'évaluer la capacité de chaque méthode en fonction de la difficulté des relations. Bien que le nombre de données soit inférieur pour ces sous-ensembles (572 images pour S1+S2\_N2-3, et 199 images pour S1+S2\_N3, avec toujours 75% pour l'apprentissage et 25% pour le test), on vérifie que la performance reste stable. On peut alors observer que l'écart entre les méthodes basées sur l'histogramme de forces et les autres augmente, puisque le score des premières reste élevé alors que les méthodes comparatives ont beaucoup plus de difficultés à appréhender les configurations plus complexes. On constate cependant que la combinaison

FIGURE 2.14 – Courbes d'apprentissage (*loss* et *accuracy*)

Entraînement-validation sur 2SimpleShapes1-2 et test sur l'image GIS, avec le modèle *SmallCNN* entraîné à partir de zéro (moyenne sur les 4 divisions du jeu de données).

$F0+F2$  est plus performante que le  $dFB$  dans ces tests, alors que le  $dFB$  comprend ces deux forces, ce qui laisse entendre que le classifieur employé n'est pas forcément le plus adapté pour le bandeau. Des tests avec un autre classifieur montrent que le  $dFB$  est capable d'atteindre les mêmes performances que  $F0+F2$  voire de les dépasser pour les configurations les plus difficiles ( $N3$ , cf. § suivant).

Des tests spécifiques sur S1 ou S2 seuls ont également été réalisés pour évaluer la difficulté de chaque jeu de données. Les performances inférieures sur S2 par rapport à S1 pour toutes les méthodes permettent de confirmer que S2 est un jeu de données plus difficile que S1, ce qui était attendu car il contient des configurations plus complexes (cf. Tableau 1.1).

Enfin, la classification sur *SpatialSense* (S3) semble beaucoup plus difficile, compte tenu des scores inférieurs obtenus sur ce jeu de données (même avec un apprentissage sur S3 lui-même), qui contient des relations spatiales dépendantes du point de vue "3D". Cela peut s'expliquer par la complexité de ce point de vue, mais aussi par des variations importantes dans les annotations (en utilisant soit le point de vue image, soit le point de vue de l'objet). Ici, les méthodes basées sur l'histogramme de forces sont comparables à la référence *bbox coords* pour les trois tests (y compris avec entraînement sur S1+S2 et avec la relation supplémentaire "dans"). Pour traiter ce point de vue 3D, il faudrait intégrer une étape de détection de pose dans la chaîne, afin de modifier le bandeau de forces en fonction de celle-ci, et également corriger les annotations qui ne sont pas cohérentes.

### Impact du modèle de classification

Comme détaillé dans la Section 2.2.4.2, deux types de modèles ont été utilisés dans cette étude : *SmallCNN* qui a été entraîné à partir de zéro, et *SqueezeNet* qui a été pré-entraîné sur IMAGENET et adapté à nos données. L'impact du modèle utilisé pour classifier les relations spatiales est évalué en comparant les résultats sur les mêmes jeux de données. Les résultats pour le modèle *SqueezeNet* et la différence avec *SmallCNN* sont donnés dans le Tableau 2.2.

Pour le  $dFB$  et sa fusion avec *bbox image*, les scores sont très similaires pour les deux modèles, sauf sur l'image SIG, où ils étaient déjà particulièrement bons avec le modèle *SmallCNN*, et pour les

Datasets	<i>dF-banner</i>			<i>bbox image</i>			<i>dFB+image</i>			<i>bbox coords</i>	
	OA	diff.	STD	OA	diff.	STD	OA	diff.	STD	OA	STD
<b>Entraînement sur S1+S2</b>											
Test sur S1+S2	<b>97.74</b>	-0.15	0.76	95.44	+2.36	1.34	<b>97.64</b>	+0.40	0.89	96.54	0.19
Test sur S1	<b>98.62</b>	-0.34	0.33	97.26	+0.93	1.02	<b>98.52</b>	-0.09	0.74	97.92	0.30
Test sur S2	<b>97.15</b>	-0.17	0.96	94.81	+3.09	1.91	<b>97.49</b>	+0.53	1.13	95.83	0.58
Test sur S1+S2_N2-3	<b>98.57</b>	+0.92	0.79	97.34	+1.85	0.79	<b>98.77</b>	+0.20	0.47	93.03	0.82
Test sur S1+S2_N3	<b>98.44</b>	+2.61	2.00	96.88	+5.21	1.21	<b>98.96</b>	+2.08	1.20	86.98	1.04
Test sur SIG	<b>96.19</b>	-3.28	1.08	89.21	+1.84	6.89	<b>96.97</b>	-0.66	1.32	<b>96.19</b>	0.26
<b>Test sur S3 (4 classes)</b>											
Entraînement sur S3	<b>71.60</b>	+0.25	2.02	69.24	+2.48	1.47	<b>72.11</b>	-0.09	1.66	71.78	4.69
Entraînement sur S1+S2	<b>67.42</b>	+1.23	1.63	61.94	-2.95	0.57	65.54	-1.25	1.29	<b>66.70</b>	1.16
<b>Test sur S3 (5 classes)</b>											
Entraînement sur S3	64.91	-0.53	0.86	65.12	+4.19	1.59	<b>68.71</b>	+1.97	0.92	64.38	1.42

TABLEAU 2.2 – Résultats de classification sur les ensembles de test avec le modèle *SqueezeNet* (taux de bonne classification maximal sur 50 *epochs* – OA, différence avec le modèle *SmallCNN* – diff., et écart-type sur 4 partitions de validation croisée – STD, tous en %). Les jeux de données S1 et S2 renvoient à *2SimpleShapes*, SIG à l’image de télédétection et S3 à *SpatialSense*, qui contient la classe supplémentaire "dans". Pour chaque test, le meilleur score est en rouge et gras, les scores proches en rouge. Les écarts positifs sont en orange et les négatifs en bleu.

configurations difficiles de *2SimpleShapes* (S1+S2 avec les sous-ensembles *N2-3* et *N3*), où à l’inverse les résultats sont meilleurs avec le modèle *SqueezeNet*, mais avec des écarts-types importants. Concernant la méthode *bbox image*, presque tous les résultats sont meilleurs avec le modèle *SqueezeNet* qu’avec le modèle *SmallCNN*, ce qui n’est pas surprenant puisque *SqueezeNet* a été pré-entraîné sur des images naturelles contenant des discontinuités similaires. Cela profite également à sa fusion avec le *dFB*, dont les scores sont améliorés dans la plupart des cas, dans une moindre mesure.

Ainsi, un plus gros modèle comme *SqueezeNet* pré-entraîné sur des images naturelles permet d’atteindre de meilleures performances sur des images binaires brutes (approche *bbox image*), mais pas sur des images de bandeaux de forces où un réseau beaucoup plus petit produit des résultats similaires, voire meilleurs. Cela signifie que le *dFB* est une représentation adaptée à la tâche, puisqu’il ne nécessite pas un trop gros réseau pour en déduire la relation spatiale.

#### 2.2.4.4 Analyse de l’emploi de différentes forces

##### Impact de la distribution des niveaux de forces

Avec le bandeau de forces, nous cherchons à avoir autant d’informations complémentaires que possible en ajoutant de nouvelles forces. Il est donc intéressant de déterminer quelle est la distribution optimale des valeurs de  $r$  : s’il vaut mieux prendre des forces positives plutôt que négatives, ou des forces proches de 0 plutôt qu’éloignées par exemple.

Des travaux antérieurs ont montré que les valeurs de 0 et 2 étaient particulièrement intéressantes, en raison de leur signification "physique", et que les valeurs négatives étaient également pertinentes (voir Section 2.2.2.2). Dans nos expériences, nous avons utilisé des forces dans une plage de valeurs allant un peu plus loin que ces dernières ( $[-2.8, 2.8]$ ), avec une distribution linéaire pour couvrir tout l’intervalle, et nous avons utilisé un assez grand nombre de forces (224), ce qui semble suffisant pour extraire des informations de toutes les forces de l’intervalle. Cependant, il est possible que certaines forces soient plus utiles que d’autres, et que plus de forces soient considérées autour d’elles plutôt que dans les autres parties.

Comme expérience supplémentaire, nous évaluons l’impact de la distribution en comparant à une distribution qui favorise les forces autour de zéro et une autre qui favorise les forces différentes de zéro, dans la plage de valeurs donnée, en considérant des fonctions continues afin d’éviter de grandes fluctuations dans le *dFB* :

$$r_{1,\lambda}(r) = r_{max} \cdot \frac{\sinh(\lambda r)}{\sinh(\lambda r_{max})} \quad (2.14) \quad r_{2,\lambda}(r) = r_{max} \cdot \text{sign}(r) \cdot \frac{\log(1 + \lambda|r|)}{\log(1 + \lambda r_{max})} \quad (2.15)$$

Nous avons pris  $\lambda = 1$  pour  $r_1$ , et  $\lambda = \frac{e^{r_{max}} - 1}{r_{max}} = \lambda_2$  pour  $r_2$ , ce qui permet d'écrire :

$$r_{2,\lambda_2}(r) = \text{sign}(r) \cdot \log\left(1 + \frac{|r|}{r_{max}}(e^{r_{max}} - 1)\right) \quad (2.16)$$

Ces fonctions sont représentées sur la Figure 2.15. Des fonctions favorisant des forces proches de plusieurs valeurs (0 et 2 par exemple) pourraient également être utilisées, en concevant des fonctions dédiées.

Les résultats de classification obtenus pour chaque distribution sont donnés dans le tableau 2.3. Ils ne montrent pas de variation significative des scores, ce qui indique que d'autres distributions peuvent être utilisées, et qu'il n'y a pas d'intérêt à privilégier des valeurs particulières ici. Cependant, comme mentionné précédemment, le nombre de niveaux de force considérés est assez important ici, ce qui permet d'obtenir des informations de différentes forces quelle que soit la distribution, tant qu'elle est continue et qu'elle couvre toute la gamme. La distribution pourrait avoir un impact si moins de forces étaient utilisées, mais alors l'aspect quasi-continu du bandeau serait perdu.

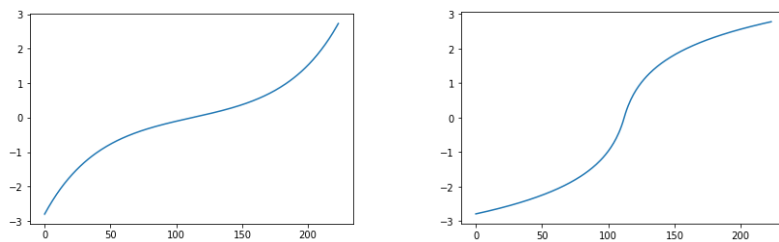


FIGURE 2.15 – Distributions utilisées pour les niveaux de forces : sinh et logarithmique

Datasets	linear		sinh			log		
	OA	STD	OA	diff.	STD	OA	diff.	STD
<b>Train on S1+S2</b>								
Test on S1+S2	97.84	0.50	97.94	+0.10	0.38	97.69	-0.15	0.53
Test on S1+S2_N2-3	97.75	0.41	96.93	-0.82	0.41	97.13	-0.62	0.47
Test on S1+S2_N3	95.83	0.00	94.79	-1.04	1.20	95.83	+0.00	0.00
Test on GIS	99.47	0.00	99.47	+0.00	0.00	99.47	+0.00	0.00

TABLEAU 2.3 – Résultats de classification sur les ensembles de test pour différentes distributions de forces (taux de bonne classification maximal sur 50 *epochs* – OA, différence avec le modèle *SmallCNN* – diff., et écart-type sur 4 partitions de validation croisée – STD, tous en %). Les écarts positifs sont en orange et les négatifs en bleu.

### Analyse par attention visuelle

Les CNN sont souvent considérés comme des boîtes noires en raison de la difficulté à expliquer comment est produite leur décision. Grâce à plusieurs travaux récents, il est maintenant possible d'identifier quelles régions de l'image sont les plus actives dans chaque couche, grâce aux cartes d'activation notamment (CAM). Par exemple, la méthode Grad-CAM [106], fournit des cartes d'activation par classe pondérées par le gradient, ce qui permet d'identifier quelles parties de l'image contribuent le plus à la décision pour chaque classe. Nous avons utilisé cette stratégie sur quelques images de bandeaux de forces pour visualiser quelles forces participent le plus à la décision. Sur de telles images, la carte de



saillance de sortie peut être considérée comme une carte d'attention où chaque ligne montre l'attention accordée à une force particulière  $r$ , tandis que chaque colonne représente l'attention liée à une direction particulière  $\theta$ .

La Figure 2.16 représente les cartes d'activation pour les échantillons de la Figure 2.9, pour la classe prédite correspondante. Étant donné qu'elles évoluent tout au long de la phase d'entraînement, nous montrons les cartes après la première époque et après la dernière époque de notre entraînement. En effet, différents intervalles de forces sont actifs selon l'époque et l'image d'entrée, montrant que le modèle se concentre sur des forces variables et que certaines forces ne sont pas pertinentes dans les premières étapes mais sont utiles pour affiner la prédiction par la suite. Par exemple, sur la première image de la Figure 2.16, les forces extrêmes sont les plus utilisées à la première époque, mais elles ne le sont plus à la dernière époque, au profit de forces plus faibles. Cela confirme que plusieurs forces sont utiles et complémentaires pour obtenir un meilleur résultat et nous donne confiance en notre approche, même si des recherches supplémentaires seraient nécessaires pour déterminer le comportement du modèle au cours de l'entraînement et l'impact de chaque force pour des configurations particulières.

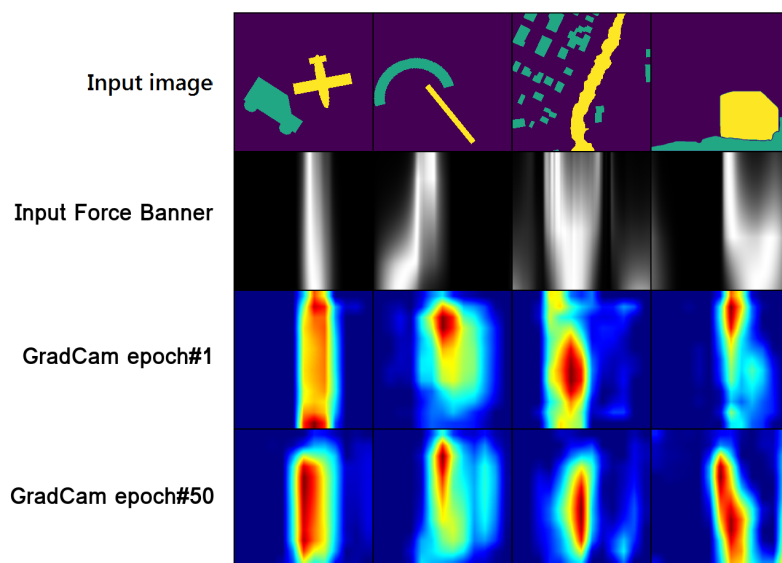


FIGURE 2.16 – Cartes d'activations obtenues avec différents bandeaux de forces, pour la classe prédite, correspondant aux images de la Figure 2.9. Les forces négatives sont en haut, les forces positives en bas.

## 2.3 Conclusion

Nous avons introduit le bandeau de forces comme descripteur de position relative entre objets binaires. Cette extension de l'histogramme de forces [79] a plusieurs avantages : (1) il hérite de l'invariance (ou quasi-invariance) aux similitudes et de la robustesse au bruit et aux petites déformations de l'histogramme des forces ; (2) il peut être calculé sur n'importe quel type d'objets et il peut décrire de nombreuses configurations différentes, des plus simples aux plus complexes ; (3) il présente l'information spatiale de façon claire et précise ; (4) il fournit plus d'informations sur la configuration qu'un seul histogramme de forces, en explorant diverses interactions entre objets, ce qui le rend plus performant sur des configurations complexes ; (5) de par sa nature planaire et son contenu proche d'une image naturelle, il peut être utilisé en entrée d'un CNN 2D classique entraîné sur ce type de données.

Ainsi, le bandeau de forces offre une représentation efficace pour caractériser la position relative entre les objets, en restituant les informations utiles pour des configurations d'objets variables. Dans cette étude, nous avons également proposé une solution pour déduire une relation spatiale en langage

naturel à partir de ce descripteur quantitatif, en utilisant des méthodes d'apprentissage pour obtenir la traduction générique. En particulier, nos expériences ont montré qu'il peut facilement être traduit en relations directionnelles et peut mieux gérer des configurations complexes que les approches par rectangle englobant pour ces relations, avec une bonne capacité de généralisation. Enfin, en raison de sa faible entropie, ce descripteur peut également être traduit en caractéristiques spatiales plus petites, afin de les utiliser dans diverses tâches de vision par ordinateur, en complément d'autres caractéristiques. Nous prévoyons d'utiliser cette méthode dans nos travaux futurs sur la classification de scènes et le suivi d'objets, en l'utilisant pour décrire une scène composée de plusieurs objets.

Ces travaux ont été valorisés dans un article de revue internationale [32], et présentés dans une version préliminaire lors d'une conférence internationale [31] et d'une conférence nationale [30]<sup>1</sup>.

---

1. <https://www.youtube.com/watch?v=gEz-dbpp38o>





## Chapitre 3

# Configuration spatiale d'un objet ou d'une scène

---

3.1	Axes de recherche de la littérature . . . . .	65
3.1.1	Généralités . . . . .	65
3.1.2	Graphes relationnels attribués (ARG) . . . . .	66
3.1.3	Sacs de relations et triplets (sujet, relation, objet) . . . . .	69
3.1.4	Vers les configurations spatio-temporelles . . . . .	70
3.2	Descripteurs par parties et comparaison par appariement . . . . .	71
3.2.1	Approche proposée . . . . .	71
3.2.2	Expérimentations . . . . .	73
3.3	Vers la reconnaissance de configurations spatio-temporelles . . . . .	75
3.4	Conclusion . . . . .	76

---

La prise en compte de la configuration spatiale des objets est un des défis majeurs de la compréhension de scènes. Plusieurs descripteurs existent déjà pour décrire des configurations de deux objets, mais le passage à des configurations de plusieurs objets reste un défi peu abordé dans la littérature. L'approche naturelle est basée sur les graphes relationnels, mais ils ont le défaut d'être difficilement comparables, nécessitant une étape d'appariement en fonction des nœuds et des arêtes, ce qui n'a pas encore été traité à notre connaissance. Nous proposons pour cela d'utiliser une description complète de l'agencement de chaque objet par rapport au regroupement de l'ensemble des autres, au moyen d'un descripteur de position relative comme l'histogramme de forces. Nous évaluons cette solution de façon qualitative sur des personnes décomposées en parties, issues de séquences du jeu d'images aériennes *Aeroscapes*, Les résultats montrent la possibilité d'utiliser une telle approche pour prendre en compte la configuration spatiale, ce que nous avons fait pour une tâche de ré-identification.

### 3.1 Axes de recherche de la littérature

#### 3.1.1 Généralités

Généralement calculés entre deux objets ou parties d'objets, les descripteurs de relations spatiales peuvent alors être associés à une décomposition de scène ou d'objet pour la ou le décrire complètement, permettant alors de reconnaître des configurations spatiales plus complexes. Ils peuvent également être intégrés dans des processus de reconnaissance de formes comme un autre type de caractéristiques, afin d'exploiter ensemble ces caractéristiques et d'obtenir une description plus complète de l'image. Néanmoins, il n'y a eu que peu de recherches sur ce sujet jusqu'à présent, et il est toujours difficile de trouver une structure pouvant contenir de nombreuses paires d'objets et qui puisse être facilement comparable

entre deux scènes, principalement en raison des problèmes d'appariement. Différentes catégorisations sont alors possibles :

1. On peut distinguer les descripteurs suivant leur niveau de décomposition de la scène (ou de l'objet) qu'ils considèrent :
  - les descripteurs globaux de la scène (ou de l'objet);
  - les descripteurs par objet (ou partie d'objet);
  - les descripteurs par relation, c'est-à-dire par paire d'objets (ou de parties d'objets).
2. On peut également distinguer les modèles qui sont ordonnés de ceux qui ne le sont pas :
  - les modèles ordonnés, où les parties sont rangées à partir de leurs attributs spatiaux, comme la position de leurs barycentres typiquement. Peu de modèles ont été proposés pour cela ;
  - les modèles non ordonnés, où les parties ou les relations sont mélangées dans le descripteur, sans que leur ordre d'apparition ait une importance ni une influence sur la traduction de la représentation, comme dans un graphe complet ou un sac de mots.

Par ailleurs, plutôt que de considérer des paires d'objets, une autre approche qui a été explorée est d'utiliser des triplets d'objets, ce qui donne des relations triangulaires. Dans cette optique, un modèle géométrique basé sur les angles formés par le triangle et nommé *Triangle Spatial Relationship* (TSR) a été introduit dans [44], puis étendu avec les  $\Delta$ -TSR [49]. Ces modèles ont notamment été utilisés pour des tâches d'indexation et de recherche d'images [92, 50].

Dans ce chapitre, nous proposons un descripteur de configuration spatiale non ordonné et par objet, comme détaillé dans la Section 3.2.1.

### 3.1.2 Graphes relationnels attribués (ARG)

Une solution naturelle pour représenter une scène décomposée en parties et en relations entre ces parties est d'utiliser un graphe, où les parties sont représentées par les nœuds (ou sommets), tandis que les relations sont portées par les arêtes, comme illustré sur la Figure 3.1. Ce type de graphe est appelé graphe relationnel attribué, ou ARG, les attributs pouvant être des étiquettes correspondant à des classes sémantiques (noms des objets, relations en langage naturel), ce qui en fait un graphe étiqueté, ou des valeurs caractérisant l'élément ou la relation (caractéristiques d'apparence, descripteurs de position relative), ce qui donne un graphe multi-valué. Puisqu'il est possible de définir une relation pour chaque paire d'éléments, il s'agit aussi d'un graphe complet, *i.e.*, toutes les arêtes existent. Pour comparer deux scènes, il est alors nécessaire de faire correspondre les sommets et les arêtes de l'un avec ceux de l'autre, en fonction de leur similarité, ce qui constitue un problème d'appariement ou d'affectation. Ce problème a notamment été étudié dans [101], pour une application de recherche de symboles complexes. Nous en donnons ici un aperçu de l'état de l'art, avant de donner un exemple d'utilisation d'ARG pour la description des relations spatiales entre les parties d'un objet.

#### Appariement de graphes

L'appariement de scènes consiste à associer les objets de l'une avec ceux de l'autre, comme illustré sur la Figure 3.2, en optimisant différents critères pour cela. Afin d'utiliser la configuration spatiale comme critère, ces scènes peuvent être représentées par des graphes relationnels, dont les nœuds représentent les objets et les arêtes représentent les relations. L'appariement de ces graphes revient à générer un autre type de liens, ou d'arêtes, entre les nœuds des deux graphes cette fois, produisant alors un graphe biparti, comme illustré sur la Figure 3.3a. Ce problème est de nature combinatoire, puisque le nombre d'appariements possibles croît de façon factorielle avec le nombre de nœuds, ce qui rend vite la recherche fastidieuse. Cependant, l'appariement d'ARG diffère des tâches d'appariement usuelles sur des graphes classiques, où l'on cherche à faire correspondre les graphes en fonction de la configuration de leurs arêtes, ce qui correspond à un problème de recherche d'isomorphisme de graphes, ou de

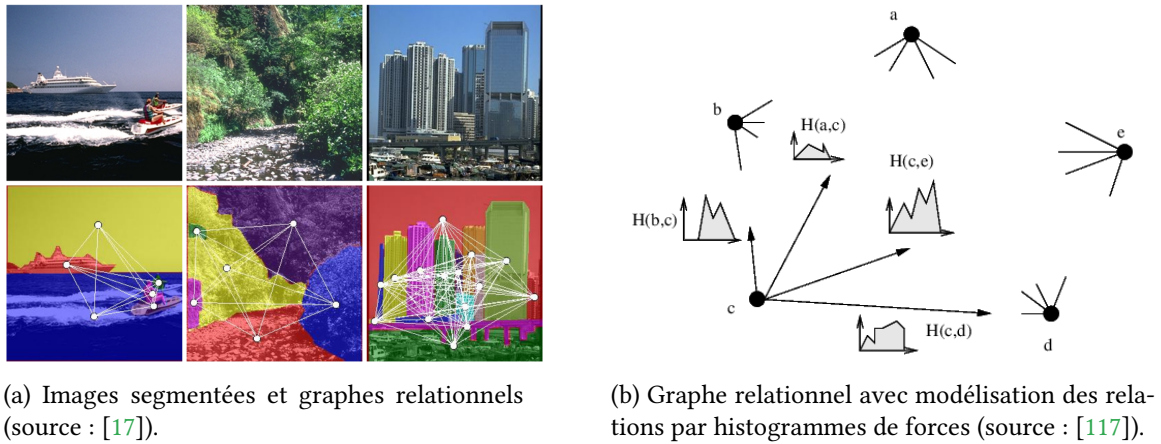


FIGURE 3.1 – Illustration de représentation de scène sous forme de graphe relationnel.

recherche de sous-graphe commun. Dans notre cas, on considère des graphes attribués complets, où le critère de correspondance n'est pas la présence d'une arête mais la similarité des attributs du sommet et des arêtes. Il s'agit donc d'une généralisation du problème de recherche d'isomorphisme à des graphes relationnels attribués, en recherchant une similarité et non une correspondance parfaite.

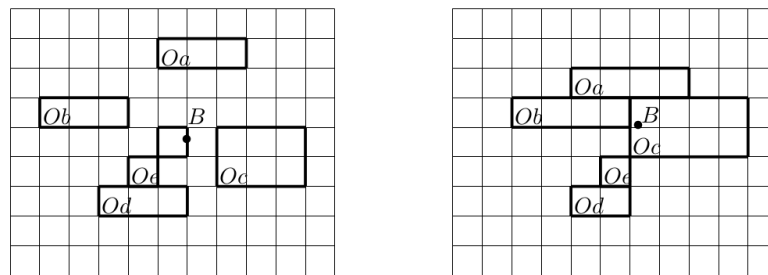


FIGURE 3.2 – Exemple d'appariement d'objets entre deux scènes.

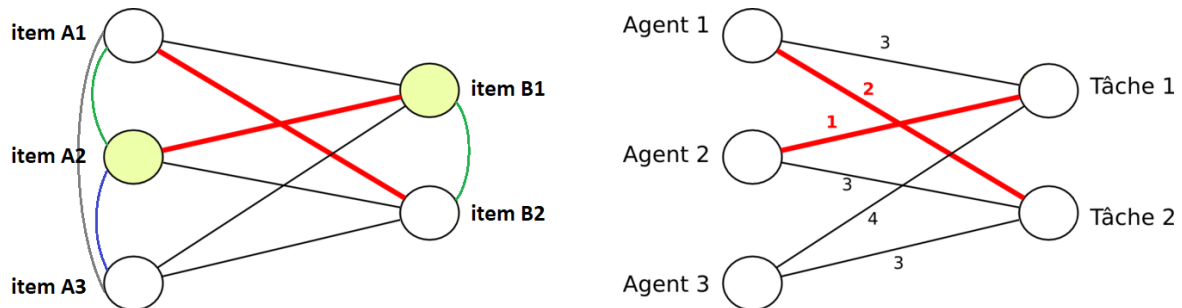
Par ailleurs, le terme d'appariement, ou couplage (*matching* en anglais) a en fait un sens plus large en théorie des graphes, puisqu'il désigne un sous-ensemble d'arêtes qui n'ont pas de sommet en commun, quel que soit le type de graphe. En fait, notre problème est plutôt un problème d'affectation ("*assignment*" en anglais), ce qui se traduit mathématiquement par la recherche de couplage maximal dans un graphe biparti, et qui est illustré sur la Figure 3.3. Dans cette tâche, le but est d'associer les éléments d'un ensemble (ou les nœuds d'un graphe) à ceux d'un autre ensemble, en associant le maximum d'éléments, un à un ou pas, et en optimisant différents critères suivant le cas. En général il s'agit de problèmes NP-difficiles, dont les solutions reposent sur des approches gloutonnes.

La version la plus courante est l'affectation linéaire, où un coût est donné pour chaque association, ce qui produit un graphe biparti pondéré, et où le but est d'optimiser (minimiser ou maximiser) la somme des poids des associations effectuées, comme illustré dans la Figure 3.3b entre des tâches et des agents. Ce problème peut être résolu en temps polynomial par l'algorithme hongrois, aussi appelé algorithme de Kuhn-Munkres, proposé par Harold Kuhn en 1955 [59] et revu par James Munkres en 1957 [82]. Cependant, cette version ne considère pas deux graphes mais deux ensembles d'éléments, sans qu'il n'y ait besoin de relation entre les éléments d'un ensemble.

Une autre version est l'affectation quadratique (QAP), où à l'inverse un poids est donné à chaque paire d'éléments des deux ensembles, ce qui en fait des graphes complets dont les arêtes sont pondérées, et où l'on cherche à associer les éléments entre les deux graphes de façon à optimiser la somme des produits des poids des arêtes ainsi associées. Ce problème est NP-difficile, mais des méthodes récentes

permettent de le résoudre assez efficacement, par exemple grâce à une approche séquentielle basée sur un apprentissage par renforcement [15, 16]. Il permet de considérer les arêtes d'un graphe et plus seulement des éléments sans connexion, mais il est limité à certains cas d'usage en utilisant comme poids les produits des valeurs. On pourrait envisager une généralisation au cas où le poids serait donné directement à chaque association d'arêtes, sans imposer ce mode de calcul, afin d'utiliser leur similarité par exemple.

Notre problème est encore différent, puisque nous considérons des graphes relationnels attribués et cherchons à associer à la fois les nœuds et les arêtes des deux graphes, en fonction de leur similarité. Il est donc une combinaison de l'affectation linéaire (pour les nœuds) et de l'affectation quadratique généralisée (pour les arêtes). Ainsi, nous faisons face à un problème qui n'a pas encore été réellement traité dans la littérature, comme en témoigne l'état de l'art disponible dans [25] par exemple. Cependant, certaines approches plus récentes l'ont abordé, comme celle de [93] qui propose d'utiliser une distance incluant à la fois une mesure de similarité sur les nœuds et sur les arêtes, avec une approche gloutonne pour l'appariement. Nous proposons quant à nous une approche permettant de se ramener à un problème d'affectation linéaire, ce qui permet d'avoir une solution assez efficace avec l'algorithme hongrois (cf. Section 3.2.1). À noter également que les graphes de relations spatiales sont orientés, puisque ces relations sont orientées, ce qui peut ajouter une complexité supplémentaire.



(a) Les éléments des graphes A et B sont appariés en faisant correspondre leurs attributs ainsi que ceux de leurs relations, représentés par des couleurs ici.

(b) Les éléments des ensembles "Agents" et "Tâches" sont appariés en minimisant le poids total des connexions entre les deux.

FIGURE 3.3 – Illustration du problème d'appariement de graphes bipartis, ou "affectation" : à gauche pour des graphes relationnels attribués, à droite sur un graphe biparti pondéré, ce qui donne un problème d'affectation linéaire (en rouge l'appariement optimal, en noir les autres appariements possibles).

### Exemples d'utilisations d'ARG

Un exemple d'utilisation d'ARG est donné avec la décomposition en histogrammes de forces, ou FHD, qui a été introduite par [41], bien qu'elle ait été utilisée avant, notamment dans les *k-formules* [118, 110]. Celle-ci consiste à décomposer un objet en parties structurales et à décrire l'ensemble des relations spatiales entre parties, ainsi que la forme de chacune des parties, grâce à un histogramme de forces pour chaque paire, comme représenté dans la Figure 3.4a. Un histogramme de forces est également calculé entre chaque partie et elle-même, ce qui permet d'obtenir un descripteur de forme de la partie. La solution a d'abord été proposée sur des images binaires, avec une décomposition par quantification des niveaux de gris. L'appariement entre deux images se fait alors naturellement en utilisant la même quantification. Puis elle a été étendue aux images couleur par [18], avec une décomposition par *clustering* basé sur la couleur, avec un appariement basé soit sur la luminosité, soit sur la forme des parties (cf. Figure 3.4b). Les histogrammes de forme et de relations sont alors utilisés pour comparer les objets grâce à une distance  $\chi^2$ , pour une tâche de recherche d'images dans [41] ou une tâche de reconnaissance dans [18]. Cette méthode de décomposition est applicable à de nombreux cas d'usage,

le *clustering* permettant en plus de mieux s'adapter aux données (voir le Chapitre 5 à ce sujet). Cependant ces premières approches reposent toujours sur un appariement des parties, ce qui peut être problématique dans certains cas.

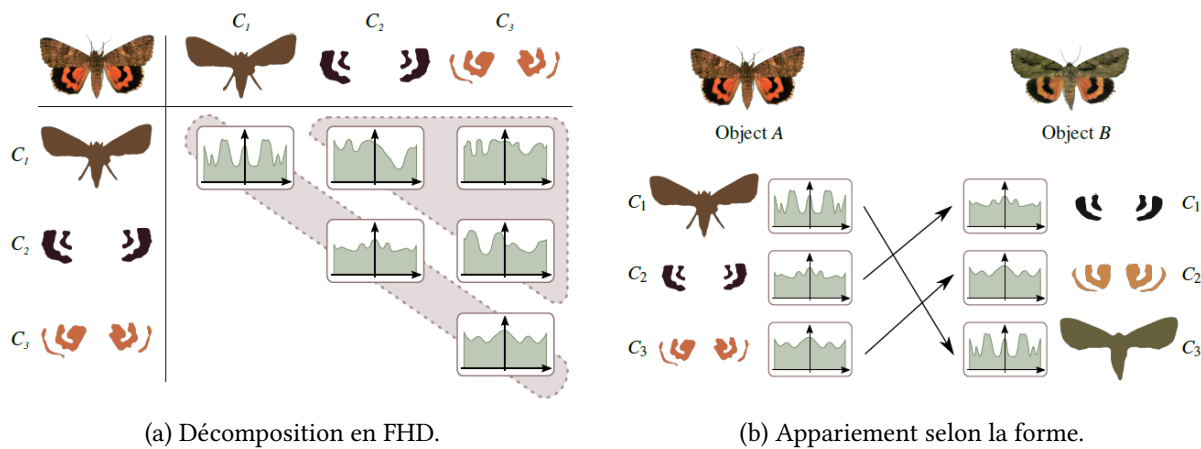


FIGURE 3.4 – Illustration de la décomposition en histogrammes de forces (FHD) pour un objet décomposé en trois parties et d'un appariement entre deux objets selon le descripteur de forme (source : [18]).

### 3.1.3 Sacs de relations et triplets (sujet, relation, objet)

Comme développé dans la Section 3.1.2, les graphes relationnels donnent une représentation naturelle et complète des constituants d'une scène ou d'un objet, mais leur comparaison se heurte au problème de l'appariement. Afin d'éviter celui-ci, d'autres approches ont été proposées avec l'idée commune de faciliter ou de ne pas nécessiter cette étape.

Tout d'abord, [102] propose de décomposer les éléments d'une scène suivant un "vocabulaire" défini de parties élémentaires, et de regrouper tous les éléments d'un même type dans un seul "objet". Cela permet d'associer directement les objets selon leur type dans la comparaison avec une autre scène, ce qui évite le problème de l'appariement. La description des relations spatiales se fait alors entre chaque paire parmi les objets ainsi générés, grâce à une signature proche de l'histogramme d'angles et appelée "Radial Line Model" (cf. Section 2.1.1.2), et la comparaison se fait en calculant la similarité entre les signatures pour chaque objet. Deux améliorations ont ensuite été proposées. Afin d'obtenir une plus grande expressivité dans la décomposition en objets, [103] suggère d'utiliser plusieurs classes différentes par type d'objet, en utilisant une étape de *clustering* pour les distinguer. Cependant la décomposition reste limitée au vocabulaire défini, qui est spécifique aux types d'images et d'objets traités. La seconde amélioration a pour but de mieux prendre en compte la configuration des objets. Pour cela, [104] propose d'ajouter à la comparaison une étape de filtrage des objets en fonction des relations topologiques entre leurs constituants. Les différentes relations présentes sont alors regroupées dans un "sac de relations", selon un ensemble de quelques relations basiques pré-définies.

En parallèle, [41] a introduit la décomposition en histogrammes de forces (FHD), qui a ensuite été étendue dans [18] (cf. Section 3.1.2). Cette décomposition et les sacs de relations ont alors été réunis dans [19, 20], utilisant directement ces descriptions pour décrire et comparer deux scènes ou objets. Pour cela, une décomposition en parties est d'abord réalisée, puis les relations spatiales et les formes sont décrites selon plusieurs prototypes obtenus par apprentissage. Chaque dimension des sacs sert à coder la présence d'un des prototypes dans une nouvelle image, comme dans les approches traditionnelles de "sacs de caractéristiques" ("*bags-of-features*"). La décomposition proposée se base sur la couleur comme dans [18], en utilisant en plus un *clustering* hiérarchique afin de générer différents niveaux de décomposition, à partir d'une segmentation initiale par *mean shift*. Utilisant l'approche du FHD, la des-

cription des relations et des formes se fait grâce à des histogrammes de forces entre les parties. Puis les prototypes pour les sacs de relations et formes sont appris grâce à un *clustering* par  $k$ -moyennes, à partir des histogrammes issus des exemples d'apprentissage. La comparaison de deux objets se fait alors en comparant directement leurs sacs de relations et de formes, sans nécessiter d'étape d'appariement des parties. Cette approche a également l'avantage d'être compatible avec les approches traditionnelles de sacs de caractéristiques, permettant des représentations hybrides qui rassemblent des informations structurelles et locales.

Des approches similaires ont été proposées pour la tâche de recherche d'image (CBIR). Par exemple, les sacs de mots d'arrangement spatial (WSA) encodent la configuration spatiale grâce à une approche par quadrants [90]. Les sacs de caractéristiques spatiaux sont quant à eux des sacs de caractéristiques classiques mais ordonnées grâce à des projections selon plusieurs axes [10]. L'approche de [50] utilise des sacs de mots combinant des descripteurs SIFT et une relation spatiale triangulaire, dans une structure de type graphe de connaissance. Et [111] utilise une description de la configuration basée sur l'histogramme de forces et un classement des configurations par "*topics*" basé sur une allocation de Dirichlet latente (LDA), afin de rechercher des configurations particulières dans une image 2D simulée (cf. Figure 1.8).

Par ailleurs, une autre approche similaire pour décrire la configuration d'une scène peut consister à la représenter par un ensemble désordonné de triplets (sujet, relation, objet), sans passer par un graphe, ce qui revient à créer un "sac de triplets". Deux options sont possibles pour décrire ces éléments : soit grâce à des descripteurs dédiés, ce qui revient à combiner les sacs de relations précédents avec les caractéristiques des objets associés, soit grâce à des triplets en langage naturel, en traduisant les descripteurs de relations (cf. Section 2.1.2.1) ou en profitant du développement des méthodes de détection directe de tels triplets (cf. Section 2.1.2.2). En apprenant des prototypes sur une base, ces sacs de triplets sont alors utilisables pour comparer des scènes de la même façon que les sacs de caractéristiques, de façon plus ou moins précise selon l'option retenue.

En particulier, l'option basée sur la détection directe de triplets a l'avantage de permettre d'utiliser des images non segmentées, sans passer par des descripteurs dédiés à la configuration spatiale. Elle permet aussi de détecter en une seule passe l'ensemble des triplets présents dans l'image, plutôt que de traiter séparément la détection des objets et la description des relations. Cependant, elle nécessite de très grandes quantités de données d'apprentissage annotées avec des triplets variés (objets et relations) pour pouvoir généraliser correctement. Plusieurs jeux de données ont été proposés récemment pour cela, comme *SpatialSense* [119] ou *Rel3D* [42] (cf. Section 1.1.3.3), mais une telle approche ne semble pas encore avoir été évaluée sur ceux-ci.

Ainsi, toutes ces représentations donnent des pistes intéressantes, mais elles ont le défaut de mélanger toutes les parties en voulant supprimer l'appariement, ce qui peut aboutir à des erreurs importantes si elles ne prennent pas en compte les caractéristiques essentielles des objets. De plus, elle demandent un volume important de données d'apprentissage afin d'obtenir une variété suffisante dans les prototypes ou les *topics*, ou de généraliser correctement pour les triplets.

### 3.1.4 Vers les configurations spatio-temporelles

De nombreux cas d'usage en vision par ordinateur ne concernent pas des prises de vues isolées que l'on souhaiterait comparer mais des acquisitions répétées voire en continu de scènes où les éléments observés évoluent, de façon plus ou moins importante. Cette dimension temporelle est alors cruciale pour la compréhension de la scène et en particulier des relations spatiales entre ses constituants. Elle demande un raisonnement particulier, ce qui est notamment introduit par [100], qui étudie l'évolution possible des relations topologiques et directionnelles à partir d'une configuration donnée. Dans la continuité des graphes relationnels, une solution pour cela est de représenter l'évolution d'une scène par un graphe spatio-temporel incluant les relations spatiales, ce qui a été formalisé par [29, 28]. Pour cela, cette approche introduit également la notion de filiation entre entités pour traduire leur identité ou



leur décomposition entre des instants différents. Une relation spatio-temporelle est alors la combinaison d'une relation de filiation temporelle et d'une relation spatiale, comme illustré sur la Figure 3.5. Cependant, ce type de représentation devient vite lourd et est donc difficilement utilisable pour des scènes qui évoluent beaucoup. Ainsi, jusqu'à présent il a surtout été utilisé pour suivre l'évolution d'entités géographiques importantes qui évoluent lentement, comme des régions d'un pays [28] ou des parcelles agricoles [65].

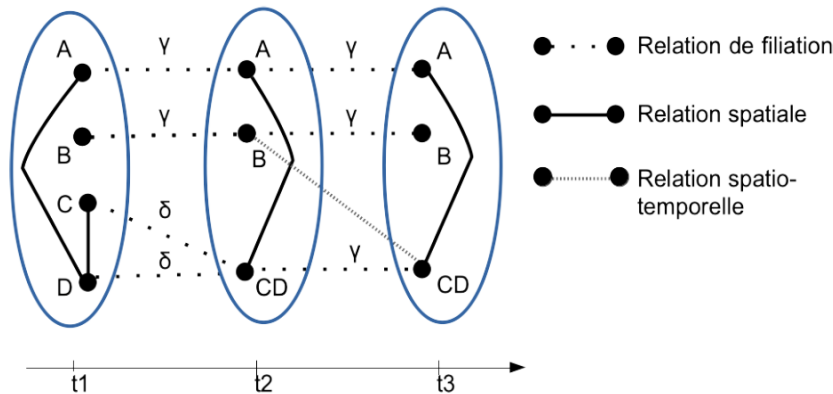


FIGURE 3.5 – Exemple de graphe spatio-temporel (source : [65], d'après [29]).

## 3.2 Descripteurs par parties et comparaison par appariement

### 3.2.1 Approche proposée

L'approche que nous proposons est une approche non ordonnée, proche de la décomposition en histogrammes de forces (FHD) proposée dans [41, 18], avec une description des parties puis un appariement grâce à ces descriptions. Cependant, ces approches utilisent uniquement des caractéristiques propres à chaque partie (leur couleur ou leur forme) pour réaliser l'appariement, considérant les relations spatiales uniquement après l'appariement, pour le calcul de la similarité. Or, la configuration spatiale des parties a un intérêt majeur également pour l'appariement, en considérant que celle-ci ne change pas au sein d'une même classe, ce qui peut ne pas être le cas pour d'autres attributs. Nous proposons ici de l'exploiter en utilisant un descripteur par parties décrivant la relation de celle-ci par rapport à l'ensemble des autres, en utilisant l'histogramme de forces comme dans le FHD. Cette solution de descripteurs par parties permet de ne pas complexifier la tâche d'appariement, qui reste alors un problème d'affectation linéaire en calculant des distances entre les descripteurs des deux objets, comme pour l'appariement sur la forme dans [18]. Cela permet d'utiliser l'algorithme hongrois pour la résolution, ce qui a aussi l'avantage de produire un score pouvant être utilisé directement comme score de similarité global. Nous détaillons ici ces différentes étapes. Par ailleurs, nous avons utilisé cette approche en combinant plusieurs types de descripteurs dans une tâche de ré-identification, ce qui est détaillé dans le Chapitre 6 de ce manuscrit.

#### Description par parties "un contre tous"

L'approche usuelle pour décrire numériquement la configuration spatiale des parties d'un objet (ou des objets d'une scène) est de décrire les relations spatiales entre ces parties prises deux à deux (et éventuellement leur forme). Considérer des paires d'objets demande alors un traitement particulier, par des graphes relationnels par exemple, ce qui est plus complexe que pour des descriptions isolées (cf. Section 3.1). Une alternative plus simple consiste à utiliser la position de chaque partie pour la décrire, ce qui permet d'avoir une représentation globale de la configuration en considérant l'ensemble des



parties. Cette position peut être simplement celle du barycentre, mais celui-ci ne permet pas de donner une description complète en ne tenant pas compte de la forme et de l'arrangement des objets, comme souligné par Rosenfeld [96]. Un ensemble de points d'intérêts ou de coins peut réduire ce problème, mais la comparaison de configurations est alors difficile à réaliser car cela ajoute une combinatoire importante, d'autant plus lorsque des transformations géométriques sont appliquées.

Il existe alors une autre solution, qui consiste à décrire la relation de chaque objet par rapport à l'union de l'ensemble des autres, ce que l'on a appelé une relation "un contre tous". Cette solution permet d'avoir un descripteur par partie et non par relation, réduisant ainsi leur nombre et la complexité de la tâche d'appariement. En effet, l'information de configuration est ici résumée dans les nœuds du graphe, ce qui évite de devoir utiliser des arêtes pour chaque paire d'objets, passant de graphes relationnels comme sur la Figure 3.3a à des ensembles de nœuds comme sur la Figure 3.3b. En valuant les liens entre deux ensembles par la similarité de leurs nœuds, c'est-à-dire de leurs relations "un contre tous", la tâche d'appariement se ramène alors à un problème d'affectation linéaire, qui peut être résolu par l'algorithme hongrois. La relation entre un objet et les autres peut être décrite par n'importe quel descripteur de relation spatiale, le choix dépendant de l'importance que l'on souhaite donner aux différents aspects de celle-ci (cf. Section 2.1.1).

Dans nos travaux, nous proposons d'utiliser l'histogramme de forces [79], qui est un descripteur complet et assez rapide à calculer avec les matériels actuels, suivant la précision choisie. Il s'agit d'un descripteur directionnel dans le sens où il décrit l'interaction entre deux objets suivant l'ensemble des directions, avec un pas d'échantillonnage donné, ce qui permet de tenir compte de la forme de chacun, voire de leur agencement suivant le paramètre de force utilisé. L'origine et la définition mathématique de ce descripteur sont détaillées dans le Chapitre 2 (Sections 2.1.1.2 et 2.2.2.1). La comparaison entre les descripteurs des parties de deux objets peut alors se faire par comparaison d'histogramme, en passant éventuellement par une étape de normalisation afin d'être robuste aux similitudes [53]. La Figure 3.6 illustre cette approche.

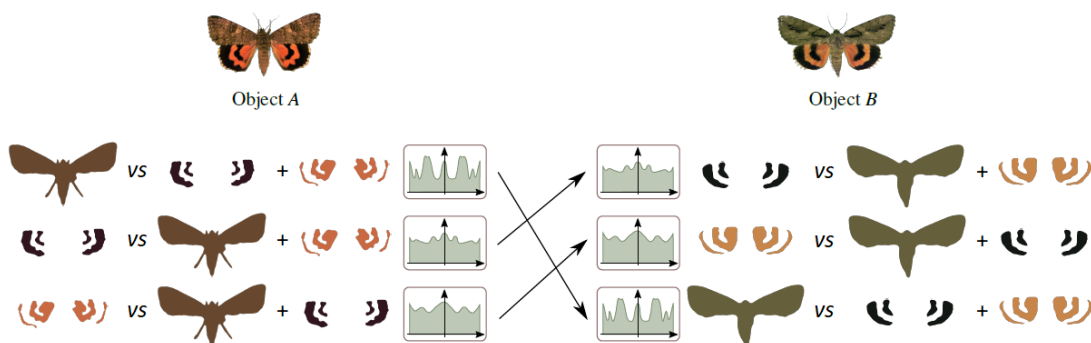


FIGURE 3.6 – Exemple d'appariement selon les relations à l'aide d'histogrammes de forces "un contre tous" (adapté de [18]).

### Appariement avec l'algorithme hongrois

Pour comparer des scènes ou des objets dont on ne connaît pas la correspondance entre les parties, il est nécessaire de passer par une étape d'appariement entre ces parties. L'objectif est d'associer chaque partie de la première scène à la partie la plus similaire dans la seconde scène, selon un critère de similarité défini, comme illustré sur la Figure 3.4b en prenant comme critère la forme des parties. Il s'agit donc d'un problème d'optimisation, où l'on cherche à minimiser la distance entre les descriptions des deux scènes.

Une solution pour cela est de définir une fonction de coût pour chaque association de parties entre les deux scènes, typiquement en utilisant la similarité entre leurs descripteurs, de façon à créer un

graphe biparti valué. Le problème de l'appariement devient alors un problème d'affectation, comme illustré sur la Figure 3.3b : il s'agit d'associer au mieux chaque élément d'un ensemble  $\mathcal{A}$  à un élément de l'ensemble  $\mathcal{B}$ , de façon à maximiser (ou minimiser) le score global de l'ensemble des affectations. Il peut être résolu en temps polynomial avec l'algorithme hongrois (en  $O(n^3)$  avec  $n$  le nombre de nœuds), dont des implémentations sont disponibles dans de nombreux environnements de développement.

Nous proposons d'utiliser cet algorithme pour l'appariement de parties entre deux scènes, en utilisant comme fonction de coût la distance (ou similarité) entre les descripteurs de ces parties. Cette méthode a plusieurs avantages : d'une part elle permet d'utiliser les descripteurs des parties pour réaliser l'appariement, et d'autre part elle produit un score global (la somme des coûts pour les affectations effectuées) qui peut être utilisé comme score de similarité entre les deux scènes. Cependant, elle a le défaut d'être limitée à des appariements un-à-un, ce qui suppose que la décomposition en parties n'a pas à être modifiée. Plusieurs méthodes sont possibles pour comparer des histogrammes de forces, et sont décrites dans la Section 2.1.1.3. Suivant les conclusions de [73], nous suggérons d'utiliser l'indice de Ružička (ou son complémentaire la distance de Soergel), qui est la généralisation de l'indice de Jaccard/Tanimoto (plus connu sous le nom de IoU) aux valeurs réelles, comme défini dans l'équation 3.1 :

$$\text{sim}(H_1, H_2) = \frac{\sum_{\theta} \min(H_1(\theta), H_2(\theta))}{\sum_{\theta} \max(H_1(\theta), H_2(\theta))} \quad (3.1)$$

### 3.2.2 Expérimentations

Afin de valider notre approche, nous avons mené plusieurs tests sur plusieurs séquences d'images issues du jeu de données *Aeroscapes* [87]<sup>1</sup>, dont nous avons extrait les personnes grâce aux masques de segmentation fournis, puis que nous avons décomposées en parties grâce à une segmentation basée sur la couleur (cf. Section 1.2.2.3 et Chapitre 5). Nous avons ensuite calculé les histogrammes de forces "un contre tous" sur ces parties, et utilisé l'algorithme hongrois pour l'appariement et le calcul du score de similarité, comme suggéré dans la Section 3.2.1. Nous avons aussi utilisé un appariement basé sur la position des barycentres à des fins de comparaison pour la description de la configuration spatiale, ainsi que sur d'autres caractéristiques (l'aire et la couleur moyenne). Ces expérimentations font partie d'une étude plus large menée sur la ré-identification, qui est détaillée dans le Chapitre 6 de ce manuscrit. Nous avons mené cette étude en langage *python*, en utilisant l'implémentation de l'algorithme hongrois de la bibliothèque *SciPy*, et en calculant la similarité entre histogrammes de forces grâce à l'indice de Ružička (indice de Jaccard/Tanimoto généralisé, cf. équation 3.1).

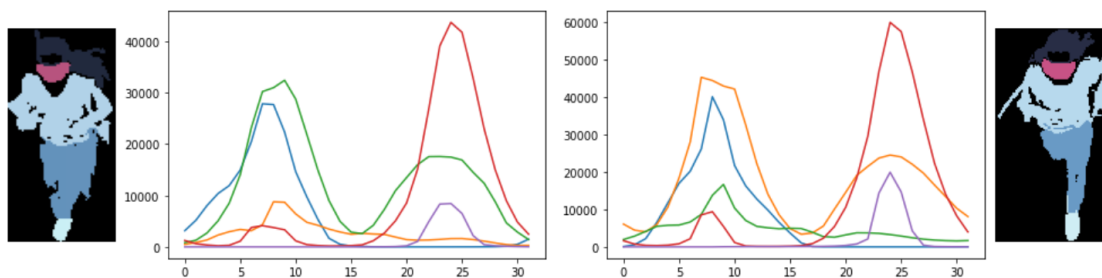


FIGURE 3.7 – Exemples de descriptions d'objets par histogrammes de forces "un contre tous", sur deux images du jeu de données *Aeroscapes*, calculés avec la force  $F_0$  et selon 32 directions (en abscisse). Chaque courbe représente la relation entre une partie et l'ensemble des autres.

La Figure 3.7 donne les histogrammes de forces "un contre tous" pour deux images de la séquence 040000, à quelques secondes d'écart, en les calculant avec la force  $F_0$  et pour 32 directions. On peut constater que l'on retrouve les mêmes formes d'histogrammes, avec des variations similaires dans

1. <https://github.com/ishann/AeroScapes>

chaque image, ce qui permet d'en déduire facilement une correspondance. Dans cet exemple simple, les parties sont correctement appariées par l'ensemble des critères utilisés, et le score de similarité est élevé quel que soit le critère.

Nous avons ensuite cherché à visualiser l'évolution de la similarité sur l'ensemble d'une séquence. Pour cela nous avons comparé chaque image à une image de référence (la première de la séquence) ou à la précédente dans la séquence. Des images représentatives extraites de cette séquence sont données dans la Figure 1.13. Dans cette séquence, la configuration et les couleurs restent globalement les mêmes, hormis au milieu (images 23 à 32) où la personne retire sa casquette et où le visage se retrouve dans la même composante que le fond, tandis que les cheveux sont séparés en deux composantes. Les Figures 3.8 et 3.9 donnent les distances totales d'appariement obtenues sur la séquence 040000, en utilisant comme critère soit les coordonnées des barycentres, soit l'histogramme de forces, soit une combinaison de ceux-ci avec l'aire et la couleur. Cette dernière peut être considérée comme la meilleure solution possible, traduisant le mieux ce qu'on attendrait comme résultat, à l'instar d'une vérité terrain. Sur l'ensemble de la séquence, on constate que l'appariement basé sur la configuration avec l'histogramme de forces est plus fiable que celui basé sur la position des barycentres ou la taille des parties (non reportée ici), avec une évolution similaire. À noter cependant qu'il s'agit du score d'appariement maximal obtenu pour le critère, et qu'il peut être inférieur à celui de l'appariement réel. Il serait alors intéressant d'annoter ces séquences avec l'appariement attendu, ou d'utiliser l'appariement obtenu avec la couleur, afin de pouvoir calculer le score pour celui-ci.

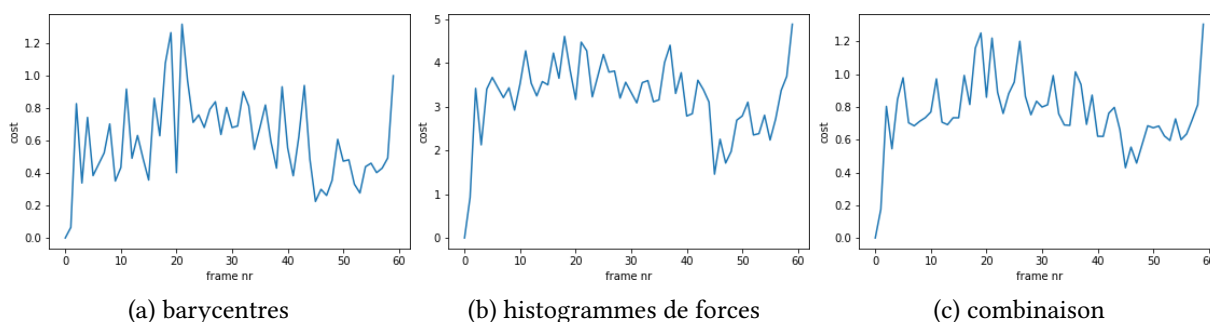


FIGURE 3.8 – Évolution de la similarité des images de la séquence 040000 du jeu de données *Aeroscapes* selon différentes caractéristiques : distance globale d'appariement avec la première image de la séquence. À noter que ces scores ne sont pas normalisés et donc pas comparables d'un critère à l'autre.

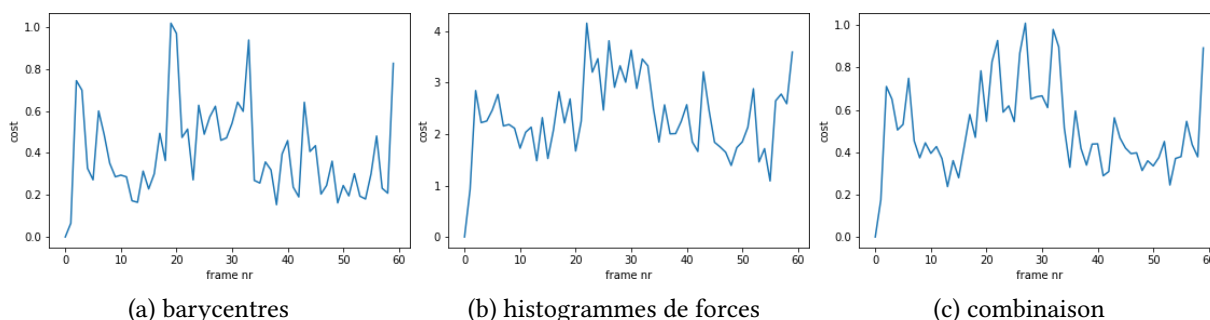


FIGURE 3.9 – Évolution de la similarité des images de la séquence 040000 du jeu de données *Aeroscapes* selon différentes caractéristiques : distance globale d'appariement avec l'image précédente dans la séquence. À noter que ces scores ne sont pas normalisés et donc pas comparables d'un critère à l'autre.

### 3.3 Vers la reconnaissance de configurations spatio-temporelles

Comme évoqué dans les Sections 1.1.3.1 et 3.1.4, la prise en compte de la dimension temporelle conduit à de nouveaux cas d'usages et à devoir utiliser des modèles adaptés, comme les graphes spatio-temporels. Une configuration spatio-temporelle pouvant être définie comme une évolution particulière des configurations spatiales, une solution pour comparer et reconnaître des configurations spatio-temporelles consiste à analyser cette évolution. Dans cette optique, nous proposons ici une étude préliminaire sur cette tâche, en étudiant l'évolution temporelle des descripteurs de configuration spatiale d'une scène simulée contenant trois objets dont un mobile. Pour cela, nous utilisons comme descripteurs les histogrammes de forces entre chaque paire d'objets, ce qui donne trois histogrammes. La séquence utilisée, qui a été générée avec *Blender*, est introduite dans la Section 1.2.2.1. Des extraits sont donnés dans la partie supérieure de la Figure 1.9. La séquence complète contient 200 images avec trois formes géométriques simples, celle du bas se déplaçant de gauche à droite.

L'évolution des trois histogrammes de forces est illustrée en représentation polaire dans la Figure 3.10, en superposant tous les histogrammes obtenus pour chaque paire, dans l'ordre chronologique. Ils sont également représentés sous forme de "bandeaux temporels" d'histogrammes de forces dans la Figure 3.11, en empilant leurs représentations linéaires. La dernière paire d'objets n'évoluant pas, les valeurs de leurs histogrammes n'évoluent pas non plus. Sur les autres graphiques, on distingue clairement l'évolution de l'orientation des objets les uns par rapport aux autres, avec un secteur angulaire d'interaction restreint au début, puis s'élargissant alors que l'objet O2 passe proche des autres, et se réduisant à nouveau.

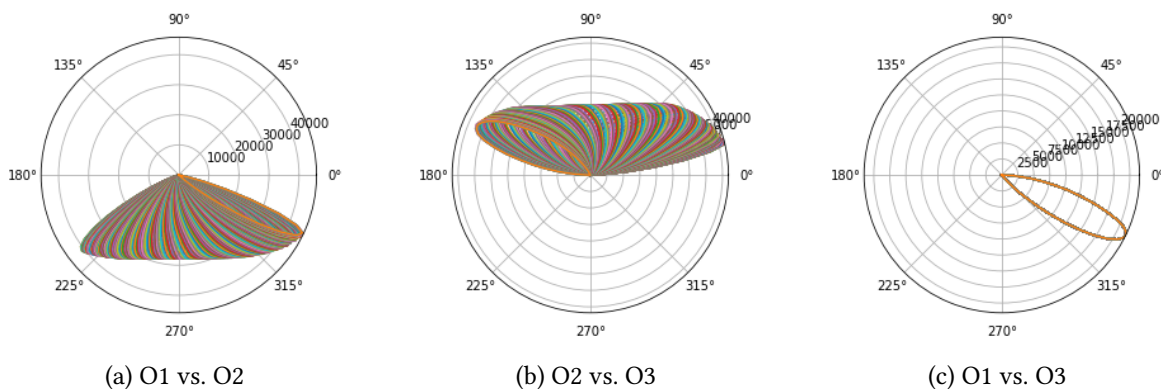


FIGURE 3.10 – Évolution des histogrammes de forces par paire d'objets sur une scène comportant trois objets, en représentation polaire. À noter que les échelles sont différentes d'un graphique à l'autre.

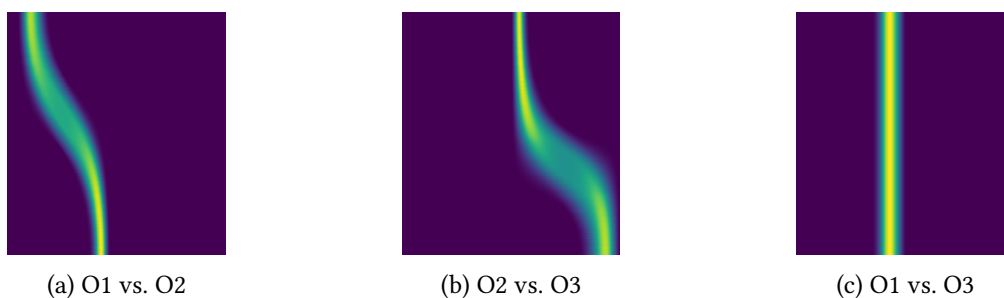


FIGURE 3.11 – Évolution des histogrammes de forces par paire d'objets sur une scène comportant trois objets, représentés par des "bandeaux temporels". Chaque ligne représente un histogramme de forces à un temps donné, sous forme de carte de chaleur (les valeurs les plus élevées sont en jaune et les plus faibles en violet), avec en abscisse la direction. L'échelle est la même pour tous les histogrammes (toutes les lignes) d'un bandeau mais elle peut être différente d'un bandeau à l'autre (pas de normalisation).

Nous nous sommes ensuite intéressés à l'évolution de la similarité de la configuration spatiale au cours de la séquence. Pour cela, nous visualisons l'évolution de la similarité entre histogrammes au cours du temps, en comparant chacun au premier histogramme de la séquence, pour chaque couple d'objets. Les résultats obtenus sont donnés dans la Figure 3.12, en utilisant l'indice de Ružička pour le calcul de la similarité (cf. équation 3.1 et Section 2.1.1.3). Une estimation de la similarité de la configuration globale peut alors être obtenue en combinant les scores de l'ensemble des paires.

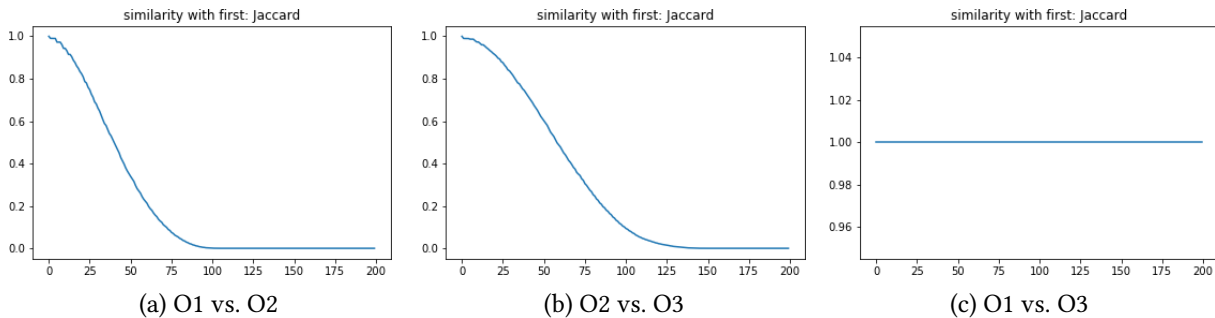


FIGURE 3.12 – Évolution temporelle de la similarité entre histogrammes de forces au cours d'une séquence, par rapport à la première image et calculée avec l'indice de Ružička, pour chaque paire d'objet dans une scène contenant trois objets. À noter que les échelles sont différentes d'un graphique à l'autre.

Ainsi, cette représentation a l'avantage d'offrir une vision simple et synthétique de la configuration spatio-temporelle. Cependant, elle suppose un appariement des objets au préalable et nécessite toujours un descripteur par paire d'objets. Une alternative est alors d'utiliser des histogrammes "un contre tous", ce qui permet de n'en avoir qu'un par objet et de les utiliser pour l'appariement (cf. Section 3.2.1). La suite de cette étude serait de comparer des scènes en fonction de leurs configurations spatio-temporelles, en utilisant cette représentation. Pour cela, il serait intéressant de générer des données dédiées à l'étude de configurations particulières, comme celles identifiées dans la Section 1.1.3.1.

### 3.4 Conclusion

La configuration spatiale est un aspect important de la perception d'une scène ou d'un objet, en plus d'autres aspects comme la forme ou la couleur des objets. Intégrant à la fois les notions de forme et de relations spatiales des parties présentes, pour des scènes ou des objets en contenant potentiellement beaucoup, sa modélisation n'est pas aisée et encore peu de travaux ont traité cette problématique. D'une part, utiliser simplement la position de points d'intérêts dans les objets (barycentres des parties, coins, blobs) n'est pas suffisant pour décrire correctement la forme et l'agencement de leurs parties. Et d'autre part, la plupart des descripteurs de configuration spatiale existants considèrent des configurations de deux objets, le passage à des modélisations en contenant davantage n'étant pas immédiat.

Plusieurs approches ont néanmoins été proposées pour cela, après une étape de décomposition en parties, comme les sacs de relations ou les graphes relationnels attribués (ARG). La première a trouvé plusieurs utilisations pour la tâche de recherche d'images, mais elle a le défaut de mélanger les parties, ce qui peut générer des confusions entre des configurations similaires. Une solution pourrait être d'utiliser des sacs combinant dans un même descripteur des caractéristiques classiques et de configuration spatiale, à l'instar des triplets (sujet, relation, objet), mais celle-ci demande beaucoup de données d'apprentissage pour bien généraliser. La seconde approche basée sur les graphes est plus naturelle, mais elle se heurte au problème de l'appariement, ou de l'affectation, pour pouvoir comparer des représentations. Un exemple est le FHD [41, 18], où les relations entre parties sont décrites par des histogrammes de forces. L'appariement entre les parties de deux objets se fait alors selon des attributs propres à ces parties (couleur et forme), sans exploiter leurs relations.

L'approche que nous proposons est similaire à celle du FHD, avec une variante afin de pouvoir prendre en compte la configuration spatiale lors de l'appariement. Après avoir obtenu une décomposition de l'objet en parties, les parties sont décrites par différents descripteurs et leur configuration par des histogrammes de forces, mais ces histogrammes sont calculés par objet, par rapport à l'union de l'ensemble des autres. Cela permet de traiter uniquement des descripteurs par partie et de calculer des distances sur ces parties entre deux acquisitions différentes, restant ainsi dans le cadre d'un problème d'affectation linéaire. Ce problème d'optimisation peut alors être résolu avec l'algorithme hongrois, évitant celui plus complexe de la comparaison de graphes relationnels. Nos expérimentations préliminaires ont montré l'intérêt de cette approche, qui doit maintenant être évaluée sur davantage de données, par exemple sur le jeu de données *Peale* utilisé par [18]. Nous avons également utilisé cette approche pour une tâche de ré-identification, qui est détaillée dans le Chapitre 6.

Enfin, nous avons étudié le passage à des descriptions de configurations spatio-temporelles, en observant l'évolution de tels descripteurs au cours d'une séquence, pour des objets ou parties déjà appariés. Ces travaux ont permis de faire émerger des représentations visuelles explicites de telles configurations, avec le bandeau temporel d'histogramme de forces notamment. La combinaison de ces travaux avec les précédents sur les configurations spatiales ou l'extension de ceux sur les graphes spatio-temporels [29] pourrait alors permettre d'aller vers des modèles utilisables pour comparer et reconnaître des configurations spatio-temporelles.



## **Partie II**

# **Segmentation pour la décomposition de scène**





## Chapitre 4

# Segmentation aidée par des annotations faibles lors de l'inférence

---

4.1	Introduction . . . . .	81
4.1.1	Contexte et positionnement des travaux . . . . .	81
4.1.2	Axes de recherche de la littérature . . . . .	83
4.1.2.1	Transfert de segmentation . . . . .	83
4.1.2.2	Segmentation à l'aide d'annotations faibles . . . . .	83
4.2	Combinaison de critères visuels et sémantiques pour la sélection de segment	84
4.2.1	Approche proposée . . . . .	84
4.2.2	Critères de sélection . . . . .	85
4.2.3	Combinaison de critères . . . . .	87
4.2.4	Expérimentations et résultats . . . . .	87
4.2.4.1	Données et protocole expérimental . . . . .	87
4.2.4.2	Résultats et discussion . . . . .	88
4.3	Conclusion . . . . .	90

---

Dans ce chapitre, nous proposons une chaîne de traitement permettant d'extraire automatiquement des objets segmentés dans une image, à partir des labels ou des rectangles englobants. Lorsqu'un label est fourni, notre système recherche le label le plus proche dans la liste des candidats, en utilisant une mesure de similarité sémantique. Et lorsqu'un rectangle englobant est fourni, il recherche le candidat avec la meilleure couverture, en fonction de plusieurs critères géométriques. Associé à un modèle de segmentation sémantique entraîné sur un jeu de données similaires, ou à un bon algorithme de proposition de régions, ce système offre une solution simple pour segmenter efficacement un jeu de données sans nécessiter d'apprentissage supplémentaire. Une étude expérimentale menée sur PASCAL VOC 2012 atteste que de tels critères permettent de sélectionner la proposition avec le meilleur score de segmentation (IoU) dans la plupart des cas, et ainsi de tirer le meilleur parti de la pré-segmentation.

### 4.1 Introduction

#### 4.1.1 Contexte et positionnement des travaux

La segmentation est une tâche incontournable dans l'analyse d'images depuis plusieurs décennies, du traitement d'images avec des solutions fondées sur la détection de contours et de régions, à l'apprentissage automatique avec des modèles (profonds) entraînés sur de grands ensembles d'images annotés. Les images segmentées sont utiles pour de nombreuses tâches de vision par ordinateur de plus haut niveau, car elles fournissent une délimitation des objets présents dans la scène, ou des parties constituant un objet. Par ailleurs, des informations pertinentes pour leur reconnaissance ou leur interprétation

(texture, forme, pose, configuration spatiale, etc.) peuvent être calculées à partir des objets segmentés. Dans nos travaux, cette tâche nous intéresse particulièrement pour pouvoir calculer des descripteurs de relations spatiales destinés aux images binaires.

Initialement, la segmentation d'images avait pour objectif de partitionner le contenu de l'image en régions homogènes, en attribuant chaque pixel à une région. Actuellement, beaucoup de travaux font référence au concept de segmentation sémantique, dont le but est de trouver le meilleur label pour la région ou le pixel. Les sorties sont appelées "segments" et sont constituées d'un label et d'un ensemble de pixels de l'image, qui peuvent être donnés sous forme de "masques de segmentation". Une variante est la segmentation d'instances, où les différentes instances de la même classe sont représentées dans différents segments.

En parallèle, la segmentation d'objets consiste à extraire un objet par rapport au fond. Des approches spécifiques existent également pour cette tâche, par exemple en partant de l'hypothèse qu'il n'y a qu'un seul objet dans l'image. D'autres hypothèses sur l'objet peuvent aussi être exploitées, comme sa position (généralement au centre), son étendue (donnée par son rectangle englobant typiquement), etc. Les solutions reposant sur le rectangle englobant et le label sont détaillées dans la Section 4.1.2.2.

La segmentation sémantique est apparue avec l'essor de l'apprentissage automatique, et plus récemment l'apprentissage profond, avec des modèles de réseaux de neurones convolutionnels (CNN) comme R-CNN [7] ou FCN [19]. Ceux-ci s'appuient sur une supervision avec des annotations denses au niveau des pixels, ce qui est fastidieux à produire sur de grands ensembles de données. Ainsi, plusieurs jeux de données ont été mis à disposition de la communauté, comme PASCAL VOC ou COCO, mais ceux-ci sont encore limités à des applications spécifiques. Par ailleurs, il a été évalué expérimentalement que "collecter des rectangles englobants autour de chaque objet de l'image est environ 15 fois plus rapide que d'étiqueter des images au niveau pixelique" [18]. Dans ce contexte, les solutions faiblement supervisées sont apparues comme une alternative moins coûteuse, n'utilisant que des annotations faibles comme des rectangles englobants ou des légendes d'images (*captions*) pour l'apprentissage. Lorsqu'elles sont disponibles à l'inférence, ces annotations peuvent être exploitées pour assister la segmentation ("segmentation à l'aide d'annotations"). Il résulte de ces différents cas des tâches de difficulté variable et des solutions spécifiques. Enfin, il est intéressant de remarquer que la segmentation est étroitement liée à la proposition de régions, qui consiste à extraire des régions cohérentes de l'image.

Bénéficiant du développement de grands jeux de données panoptiques, il est désormais possible de trouver de nombreux modèles de segmentation pré-entraînés et de les utiliser directement sur de nouvelles données avec de bonnes performances, pour de nombreuses applications. La sortie peut être une seule image segmentée avec différents labels, ou un masque de segmentation par classe avec recouvrement possible entre les classes. Dans les deux cas, il peut être utile d'extraire automatiquement un objet particulier parmi tous les segments produits, à partir d'indices à son sujet, de manière à intégrer cette étape dans un processus plus large. C'est ce problème que nous considérons ici. La segmentation faiblement supervisée est une extension naturelle de celui-ci, une solution simple consistant à combiner proposition de région et sélection, de façon à générer une supervision au niveau des pixels à partir d'images faiblement annotées.

Dans ce chapitre, nous proposons des critères complémentaires pour extraire une sortie particulière parmi plusieurs propositions, en fonction de son rectangle englobant et/ou de son label (cf. Section 4.2). Pour le rectangle englobant, nous utilisons une combinaison de critères géométriques sur la couverture de celui-ci par chaque proposition de segmentation. Et pour le label, nous utilisons la sémantique pour trouver le label proposé le plus proche de celui recherché. Nous proposons d'utiliser cette sélection en aval d'un modèle de segmentation générant les propositions, ce qui permet d'obtenir une chaîne de traitement complète pour segmenter des ensembles de données contenant des annotations faibles. De plus, cette chaîne peut être facilement intégrée comme première étape de certains modèles de segmentation faiblement supervisée, de manière à segmenter de nouvelles données sans aucune annotation disponible à l'étape d'inférence, en utilisant les annotations faibles pour l'apprentissage.

## 4.1.2 Axes de recherche de la littérature

### 4.1.2.1 Transfert de segmentation

Les modèles modernes de segmentation, quel que soit leur niveau de supervision, sont entraînés sur des données représentatives, les rendant spécifiquement adaptés à ces dernières, avec une capacité de généralisation variable. Utiliser un tel modèle pour un autre jeu de données ou une autre tâche est possible, mais cela nécessite quelques adaptations pour le rendre efficace, en exploitant les connaissances disponibles sur le nouveau jeu de test. Cette solution générale appelée "apprentissage par transfert" est déclinée en "adaptation de domaine" lorsque la tâche est modifiée, et "*fine-tuning*" lorsque le modèle est uniquement adapté à de nouvelles classes (en modifiant la dernière couche typiquement) et ré-entraîné sur les nouvelles données. Le *fine-tuning* est très utilisé pour la détection et la reconnaissance d'objets, mais beaucoup moins pour la segmentation sémantique qui requiert des annotations denses pour l'entraînement des modèles.

Dans ce contexte, [10] propose d'utiliser l'apprentissage par transfert pour la segmentation sémantique faiblement supervisée, en transférant les connaissances des catégories avec des annotations fortes vers des catégories inconnues avec des annotations faibles, grâce à une architecture encodeur-décodeur associée à un modèle d'attention visuelle. [23] propose une solution pour bénéficier à la fois de données réelles et synthétiques, en utilisant avec le réseau de segmentation un deuxième réseau dédié à l'apprentissage de la similarité des pixels synthétiques avec les pixels réels. Enfin, [21] propose d'utiliser la similarité sémantique, en repartant aussi des poids de la dernière couche du réseau d'origine, lorsque les labels sont sémantiquement proches. Il n'est évalué que sur une tâche de classification, mais l'idée peut également être utilisée pour la segmentation. Nous considérons cette approche ici, mais uniquement comme un post-traitement étant donné que nous n'avons pas accès aux annotations au niveau des pixels pour ré-entraîner le modèle.

### 4.1.2.2 Segmentation à l'aide d'annotations faibles

Nous n'utilisons pas l'expression de "segmentation faiblement supervisée" pour cette tâche car elle se réfère généralement à l'étape d'entraînement du modèle uniquement, avec des solutions basées sur un apprentissage faiblement supervisé, alors qu'ici nous considérons que les annotations faibles sont disponibles à l'inférence. En fait, il s'agit souvent de la première étape de solutions de segmentation faiblement supervisées, exploitant des annotations faibles pour générer une vérité terrain au niveau du pixel, qui est ensuite utilisée pour entraîner un modèle de segmentation classique, de la même manière que les méthodes auto-supervisées.

Deux principales catégories d'approches coexistent : celles exploitant le rectangle englobant de l'objet et celles exploitant des labels donnés au niveau image (*captions*). Typiquement, ces indices proviennent d'une première étape de détection ou de sous-titrage (*captioning*), ou bien d'annotations manuelles. Quelques solutions exploitent une combinaison de labels au niveau image et de rectangles englobants : [20] utilise l'un ou l'autre en fonction de ce qui est disponible, tandis que [17] utilise des rectangles englobants pour les objets d'intérêt et des labels au niveau image pour l'arrière-plan (*stuff objects*). Mais à notre connaissance et étonnamment, aucune solution n'exploite la combinaison d'un rectangle englobant directement avec le label de l'objet concerné.

### Segmentation d'objet à l'aide d'un rectangle englobant

Connaitre le rectangle englobant d'un objet est évidemment un indice pertinent pour en déduire son étendue spatiale réelle, d'autant plus si celui-ci est limité à l'objet ciblé. Le cas le plus simple est la segmentation objet/fond, où l'objet est seul dans l'image ou le rectangle englobant. Mais dans de nombreux cas, plusieurs objets peuvent être présents et choisir le plus pertinent peut être délicat. Alors que les premières solutions reposaient sur une interaction plus importante avec l'utilisateur, comme de

la sélection de points ou des griffonnages à l'intérieur et à l'extérieur de l'objet, GrabCut [22] a été la première solution à ne s'appuyer que sur un rectangle englobant. Cette méthode bien connue se fonde sur deux modèles d'apparence (arrière et premier plans) et sur une approche itérative par *graph-cut* pour obtenir la segmentation finale. Des variantes comme [16] exploitent l'hypothèse d'étroitesse du rectangle englobant, en ajoutant des contraintes dans la minimisation de l'énergie.

Une autre solution courante consiste à générer plusieurs propositions de segments, avec une pré-segmentation ou un algorithme de proposition de régions, et de sélectionner ou de générer la sortie finale grâce à un vote ou une combinaison. Dans la littérature, la plupart des solutions reposent sur la proposition de régions, en particulier dans le cadre faiblement supervisé où elle est utilisée comme première étape pour générer une vérité terrain au niveau pixellique. Par exemple, BoxSup [4] repose sur l'approche MCG [2] et évalue plusieurs autres solutions. [13] est également basé sur MCG mais le combine avec GrabCut, et permet en plus de gérer plusieurs instances d'un même objet. Dans notre solution, nous suggérons d'utiliser une pré-segmentation avec un véritable modèle de segmentation entraîné sur un jeu de données similaire, en utilisant par exemple le modèle Mask R-CNN [9].

D'autres pistes ont également été explorées, notamment pour la segmentation faiblement supervisée. [20] propose d'utiliser un champ aléatoire conditionnel (CRF) et de le contraindre à considérer la zone centrale du rectangle englobant comme premier plan et les pixels à l'extérieur du rectangle comme arrière-plan. [11] utilise quant à lui le modèle MIL pour gérer plusieurs instances, en y intégrant l'hypothèse d'étroitesse du rectangle englobant. Et plus récemment encore, BB-UNet [12], qui est basé sur U-Net, exploite des hypothèses sur la forme en introduisant une nouvelle couche de convolution, pour segmenter des images médicales.

## Segmentation d'image à l'aide de labels

De nombreuses solutions exploitant les labels au niveau de l'image existent dans la littérature, en particulier pour la segmentation faiblement supervisée à nouveau [14, 27, 1, 24]. Ces solutions sont généralement basées sur des cartes d'activation par classe (CAM) et un modèle d'attention visuelle pour générer la vérité terrain au niveau des pixels. L'approche de [8] peut également être mentionnée ici, puisque son objectif est de segmenter un jeu de données complet contenant des annotations faibles comme nous (ImageNet dans leur cas), mais avec une idée totalement différente. Celle-ci repose sur une approche gloutonne originale, en segmentant progressivement les objets dont les labels sont sémantiquement proches de ceux déjà segmentés, avec un modèle d'apparence pour les pixels de premier plan et un autre pour l'arrière-plan. Le modèle est initialisé avec les annotations niveau pixels de PASCAL VOC, et exploite également les annotations de rectangles englobants lorsqu'elles sont disponibles.

## 4.2 Combinaison de critères visuels et sémantiques pour la sélection de segment

### 4.2.1 Approche proposée

Nous proposons une solution à la tâche de segmentation d'objets à l'aide d'un rectangle englobant et/ou d'un label, sur une image contenant plusieurs objets et pas seulement un arrière-plan. Elle repose sur une chaîne de traitement en deux étapes appelée SegMyO (*Segment My Object*) : tout d'abord un ensemble de régions est extrait avec une solution classique de segmentation d'images, ou avec un algorithme de proposition de régions, puis la meilleure sortie est sélectionnée pour le rectangle englobant donné, et le label dans le cas de la segmentation sémantique. Concernant l'étape de segmentation, l'utilisation d'un modèle pré-entraîné suppose que ce dernier a été entraîné sur des objets et des labels similaires à ceux d'intérêt. Pour cela, il existe maintenant des modèles de segmentation performants pré-entraînés sur des ensembles de données importants et variés, pour de nombreuses applications.

Afin d'étendre le modèle appris à d'autres labels similaires, nous proposons en plus d'utiliser la similarité sémantique pour rechercher un label similaire parmi ceux du modèle. Cela peut aussi être vu comme un transfert d'un modèle entraîné sur un ensemble de labels vers un autre ensemble ne contenant pas exactement les mêmes labels, comme cela est fait dans [21].

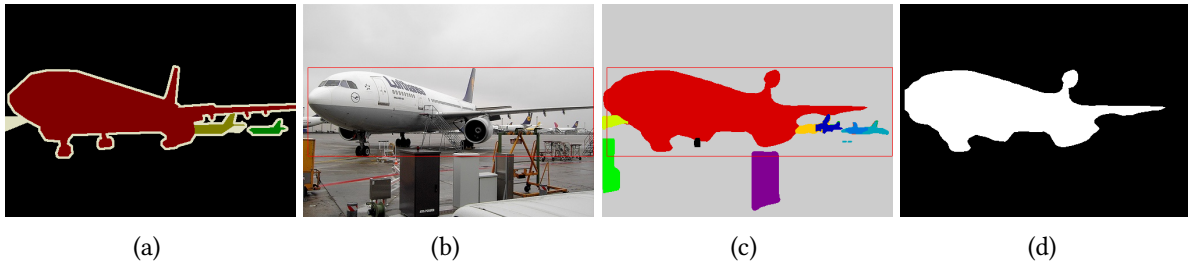


FIGURE 4.1 – Segmentation d'une image issue du jeu de données PASCAL VOC 2012 : (a) Vérité terrain (3 instances de la classe "aeroplane"), (b) image d'entrée et rectangle englobant avec le label "aeroplane", (c) masques produits par Mask R-CNN (limités au rectangle englobant - 12 objets), (d) masque sélectionné par SegMyO (de la classe "airplane").

Notre système prend en entrée :

- une image  $I$  de dimension  $W \times H$ , avec  $W$  la largeur et  $H$  la hauteur de l'image ;
- un label  $l \in L^{target}$  d'un objet présent dans l'image (e.g.,  $L^{target} = \{\text{Man, Car, ...}\}$ );
- un rectangle englobant  $B$  autour de cet objet, défini par les coordonnées de son coin supérieur gauche  $(x_1, y_1)$  et de son coin inférieur droit  $(x_2, y_2)$ , avec  $(x_1, y_1)$  et  $(x_2, y_2) \in [1, W] \times [1, H]$  ;
- une segmentation de l'image  $I$  constituée d'un ensemble de régions  $\{R_i\}_{i=1..N}$ , avec recouvrement possible, chacune détectée avec un score de confiance  $r_i$  et potentiellement accompagnée d'un label  $l_i \in L^{init}$  (e.g.,  $L^{init} = \{\text{Person, Automobile, ...}\}$ ).

Pour chaque région candidate  $R_i$ , un score est calculé à partir de la couverture du rectangle englobant, et de la similarité sémantique entre le label attendu  $l$  et le label prédit  $l_i$  si on utilise la segmentation sémantique, grâce aux critères définis dans la Section 4.2.2. Ensuite, le système retient comme segmentation de l'objet la région avec le meilleur score parmi toutes les régions candidates, comme illustré sur la Figure 4.1.

## 4.2.2 Critères de sélection

### Critères géométriques

Tout d'abord, deux critères visant à exploiter le fait que le rectangle englobant est limité à l'objet sont considérés, étant donné que ce sont des critères nécessaires et universels. Ils reposent sur l'hypothèse que le rectangle englobant contient tout l'objet et pas plus d'autres éléments que nécessaire. Ainsi, plus le rectangle englobant est précis, meilleure sera la segmentation. Cette hypothèse est utilisée dans [16, 11] et peut également être dérivée des hypothèses sur l'étendue de l'arrière-plan et de l'objet de [13]. Les critères proposés sont les suivants :

- $c_1$ . la distance relative maximale de la région aux bords du rectangle englobant, qui doit être proche de zéro pour les 4 côtés ;
- $c_2$ . l'étendue relative de la région, c'est-à-dire la partie de la région qui se trouve dans le rectangle englobant, en supposant que tout l'objet doit être dans celui-ci.

D'autres critères optionnels peuvent être définis en fonction des connaissances a priori sur la forme des objets, comme :

- $c_3$ . la couverture du rectangle englobant par l'objet, qui vise à donner plus de poids aux grands objets (c'est ce critère qui est utilisé dans BoxSup [4]);
- $c_4$ . la présence au centre du rectangle englobant, pour les objets "barycentriques" ou pour éviter les objets présents uniquement sur les bords.

Dans nos tests nous utilisons les critères  $c_1$ ,  $c_2$  et  $c_3$ . Ils sont définis comme suit dans l'intervalle  $[0, 1]$ , avec  $R_{i,B} = R_i \cap B$  la restriction de la région  $R_i$  au rectangle englobant  $B$  (cf. Figure 4.2) :

$$c_1(R_i, B) = 1 - \max \left[ \frac{d_x(R_{i,B}, B)}{W}, \frac{d_y(R_{i,B}, B)}{H} \right] \quad (4.1)$$

$$c_2(R_i, B) = \frac{\text{area}(R_{i,B})}{\text{area}(R_i)} \quad (4.2) \quad c_3(R_i, B) = \frac{\text{area}(R_{i,B})}{\text{area}(B)} \quad (4.3)$$

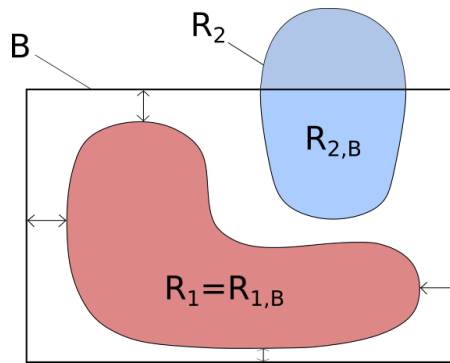


FIGURE 4.2 – Illustration des mesures géométriques utilisées pour le calcul des critères, pour deux régions candidates dans le même rectangle englobant.

### Critère sémantique

Lors de l'utilisation de la segmentation sémantique, chaque segment se voit attribuer un label à partir de l'ensemble de labels appris par le modèle. Il est donc utile de connaître celui de l'élément recherché, cependant il peut ne pas faire partie de cet ensemble. En effet, un problème majeur réside dans la spécificité et le faible nombre de classes dans la plupart des bases de données académiques usuelles. Les modèles de segmentation pré-entraînés sur ces bases de données sont donc limités à la segmentation des classes d'objets existantes dans ces bases de données. Par exemple, si la classe "chat" est apprise dans le jeu de données initial, il sera difficile de fournir comme label d'entrée "lynx" dans le jeu de données cible. Pourtant, le modèle de segmentation du chat peut être utilisé pour segmenter un lynx puisque ces objets sont visuellement proches et sémantiquement liés. C'est pourquoi nous proposons d'utiliser la correspondance sémantique entre labels pour étendre le vocabulaire aux labels en dehors de l'ensemble initial.

Pour cela, nous utilisons des outils de traitement automatique des langues et d'ingénierie des connaissances, permettant de calculer une similarité sémantique entre le label prédit et le label attendu :

$$c_{sem}(l_i, l) = \text{semantic\_similarity}(l_i, l) \quad (4.4)$$

Deux approches principales peuvent être envisagées : utiliser la similarité entre mots dans une taxonomie comme *WordNet*, ou la similarité entre des représentations de ces mots (*word embeddings*) apprises sur un corpus, avec un modèle comme *Word2Vec*. Cependant, cela nécessite que les deux labels soient présents dans le vocabulaire, ce qui nécessite un contrôle lors du choix du modèle, ou d'effectuer des corrections manuelles.



Plusieurs méthodes sont disponibles pour calculer la similarité sémantique dans une taxonomie, comme les similarités de Rada, Resnik, Li, etc. La plupart d’entre elles sont basées soit sur des mesures structurelles entre les concepts de la taxonomie (par exemple, la longueur et la profondeur du chemin), soit sur le contenu d’information (IC), qui mesure le niveau de détail d’un concept ( $IC(c) = -\log(p(c))$  avec  $p(c)$  la probabilité d’appartenance au concept  $c$ ). Nous suggérons d’utiliser la similarité *wpath* [28], qui combine les deux approches en utilisant le contenu d’information du plus petit ancêtre commun pour pondérer la longueur de chemin la plus courte entre les concepts, et donne les meilleures performances en général. Cette mesure est définie par :

$$c_{sem,wpath}(l_i, l) = \left(1 + length(l_i, l) * k^{IC(lcs(l_i, l))}\right)^{-1} \quad (4.5)$$

avec *lcs* le plus petit ancêtre commun (*least common subsumer*), et  $k$  un paramètre définissant la contribution du contenu d’information de celui-ci.

### 4.2.3 Combinaison de critères

Un critère global peut être calculé à partir de l’ensemble des critères précédents et du score de confiance de segmentation  $r_i$ . Nous proposons ici de le calculer simplement avec une somme pondérée des différents scores, en fixant nous-mêmes les poids de chaque critère. D’autres solutions sont possibles, par exemple en apprenant les poids, cependant nous avons constaté expérimentalement que la combinaison proposée était satisfaisante (cf. Section 4.2.4).

Comme mentionné précédemment, les critères  $c_1$  et  $c_2$  sont particulièrement importants, nous suggérons donc de leur donner un poids plus important, alors que le critère  $c_3$  est plus litigieux et devrait donc avoir un poids plus petit, à moins d’avoir des connaissances a priori sur la forme des objets recherchés. Le critère sémantique est également particulièrement significatif, nous suggérons donc de lui affecter un poids important.

Dans nos expérimentations, nous utilisons le score suivant :

$$score_1 = \begin{cases} (r_i + 2 * c_1 + 2 * c_2 + c_3 + 4 * c_{sem})/10 & \text{quand } l_i \in L^{init} \text{ et } l \in L^{target} \\ (r_i + 2 * c_1 + 2 * c_2 + c_3)/6 & \\ \text{sinon} & \end{cases} \quad (4.6)$$

Pour aller encore plus loin, les critères  $c_1$  et  $c_2$  étant des conditions nécessaires, nous suggérons de ne conserver que la valeur minimale de  $c_1$ ,  $c_2$  et le score global :

$$score_2 = \min(c_1, c_2, score_1) \quad (4.7)$$

Il est à noter que tous ces critères peuvent être impactés si plusieurs instances de l’objet sont présentes dans l’image et que la segmentation ne parvient pas à les séparer. Ils doivent donc être utilisés uniquement avec une segmentation d’instances ou avec des ensembles de données comprenant une seule instance de chaque objet.

## 4.2.4 Expérimentations et résultats

### 4.2.4.1 Données et protocole expérimental

Nous évaluons notre chaîne de traitement sur le jeu de données de segmentation PASCAL VOC 2012, en utilisant le sous-ensemble de validation (1 449 images, 3 427 objets, 20 classes). Pour générer les régions candidates, nous utilisons le modèle *torchvision* Mask R-CNN ResNet-50 FPN<sup>1</sup> [9], entraîné

1. <https://pytorch.org/docs/stable/torchvision/models.html>



sur COCO 2017 (80 classes, dont les 20 classes de PASCAL VOC mais avec des noms pouvant être différents). Le résultat de la pré-segmentation est une liste de propositions d'objets composées d'un label, d'un score et d'un masque de segmentation avec des valeurs dans  $[0, 1]$ . Nous ne conservons que les propositions dont le score de segmentation est supérieur à un seuil (fixé à 0,25). L'utilisabilité d'une solution de proposition de régions par groupement combinatoire multi-échelle (MCG) [2] est aussi évaluée, à partir des propositions précalculées disponibles en ligne <sup>2</sup>.

Les critères décrits précédemment ainsi que leur combinaison sont ensuite calculés. Pour la similarité sémantique, nous avons utilisé la taxonomie *WordNet* et la méthode *wpath* [28], avec l'environnement *python sematch* <sup>3</sup>. Comme métrique d'évaluation de la segmentation, le score d'intersection sur l'union (IoU) est calculé pour chaque objet en utilisant les annotations de segmentation d'instance fournies (après une binarisation de l'image, avec un seuil à 0,3). La moyenne sur l'ensemble des objets est donnée dans le Tableau 4.1. Afin de comparer avec d'autres méthodes, nous calculons également le score (IoU) moyen pour chaque classe d'objets, après avoir transformé nos sorties pour se conformer à la tâche de segmentation d'image, avec une seule classe par pixel. Pour ce faire, nous avons ajouté une étape de fusion des segments sélectionnés individuellement, en les insérant dans l'image de sortie du plus grand au plus petit, de façon à gérer le cas des segments qui se chevauchent. Nous nous comparons également à GrabCut [22], en tant que solution non supervisée de l'état de l'art pour notre tâche.

#### 4.2.4.2 Résultats et discussion

Les résultats de nos expériences sur la segmentation à l'aide d'annotations faibles sont indiqués dans le Tableau 4.1, pour plusieurs solutions et plusieurs critères de sélection, tandis que le Tableau 4.2 fournit une comparaison avec plusieurs modèles récents de segmentation faiblement et entièrement supervisés. Pour chaque solution, nous indiquons à la fois le niveau d'annotations utilisé pour entraîner le modèle (supervision) et le niveau d'annotations disponibles à l'inférence. Il peut s'agir de "pixels" lorsque les annotations au niveau des pixels sont utilisées, de "légende" lorsque des labels au niveau de l'image sont considérés et de "b.box" ou "b.box + label" lorsque des annotations au niveau du rectangle englobant sont utilisées.

Dans nos tests (Tableau 4.1), nous distinguons les solutions totalement non supervisées, c'est-à-dire n'utilisant aucune donnée pour construire le modèle, et celles n'utilisant pas les données d'apprentissage de PASCAL VOC mais un jeu de données similaire (ici COCO), dans un mode d'apprentissage par transfert. Comme méthode comparative naïve pour les annotations de rectangles englobants, une solution basée sur un remplissage partiel (à 90%) du rectangle est également évaluée (solution "rectangle englobant"). Enfin, concernant les solutions basées sur notre chaîne de traitement avec une pré-segmentation et une sélection de sortie, une borne supérieure pour le modèle de pré-segmentation choisi peut également être obtenue, en utilisant pour la sélection de sortie le score IoU comme critère au lieu du nôtre (ce qui suppose d'avoir accès aux annotations au niveau pixelique, donc à la vérité terrain).

Ces résultats montrent qu'une segmentation totalement non supervisée à l'aide d'annotations faibles ne donne pas de résultats satisfaisants, et ne permet pas de segmenter efficacement le jeu de données malgré l'information du rectangle englobant. En revanche, l'utilisation d'un modèle pré-entraîné sur un jeu de données similaire permet d'atteindre une bonne performance, se classant entre les modèles faiblement supervisés et entièrement supervisés, sans nécessiter d'entraînement sur le nouveau jeu de données. Cela montre qu'il est possible d'obtenir une segmentation correcte sans entraînement spécifique, grâce à un usage approprié des annotations faibles lors de l'inférence.

Concernant les différents critères proposés, ceux-ci donnent des résultats variables lorsqu'ils sont pris individuellement, avec un bon comportement des critères  $c_3$  et  $c_1$ , et dans une moindre mesure du critère sémantique (qui montre un écart important entre les deux modes de calcul du mIoU). Combiner

2. <https://www2.eecs.berkeley.edu/Research/Projects/CS/vision/grouping/mcg/>

3. <https://github.com/gsi-upm/sematch>

méthode	supervision	annotation à l'inférence	mIoU	mIoU <sub>obj</sub>
<b>non supervisée</b>				
rectangle englobant		b.box	56.50%	<b>53.16%</b>
GrabCut [22]		b.box	<b>58.82%</b>	44.29%
MCG + SegMyO <sub>C3</sub>		b.box	47.64%	44.39%
MCG + SegMyO <sub>mIoU</sub>		pixels (VT)	51.08%	47.50%
<b>transfert d'apprentissage (depuis COCO)</b>				
Mask R-CNN + SegMyO <sub>mIoU</sub>	COCO	pixels (VT)	<b>74.68%</b>	<b>70.67%</b>
Mask R-CNN + SegMyO <sub>C1</sub>	COCO	b.box	70.55%	65.04%
Mask R-CNN + SegMyO <sub>C2</sub>	COCO	b.box	53.19%	51.95%
Mask R-CNN + SegMyO <sub>C3</sub>	COCO	b.box	<b>71.02%</b>	<b>66.85%</b>
Mask R-CNN + SegMyO <sub>sem</sub>	COCO	caption	68.71%	58.19%
Mask R-CNN + SegMyO <sub>c3*sem</sub>	COCO	b.box+label	73.16%	68.63%
Mask R-CNN + SegMyO <sub>score1</sub>	COCO	b.box	73.08%	68.97%
Mask R-CNN + SegMyO <sub>score2</sub>	COCO	b.box	73.04%	68.74%
Mask R-CNN + SegMyO <sub>score1</sub>	COCO	b.box+label	<b>73.62%</b>	<b>69.35%</b>
Mask R-CNN + SegMyO <sub>score2</sub>	COCO	b.box+label	73.30%	68.93%
Mask R-CNN + SegMyO <sub>score1</sub> / rectangle englobant	COCO	b.box+label	<b>73.99%</b>	<b>69.88%</b>

TABLEAU 4.1 – Scores de segmentation sur l'ensemble de validation de PASCAL VOC, en utilisant des annotations lors de l'inférence, pour des modèles non supervisés ou entraînés sur un autre ensemble de données (Mask R-CNN entraîné sur COCO) (mIoU : moyenne sur les 21 classes, mIoU<sub>obj</sub> : moyenne sur l'ensemble des objets sans la classe d'arrière-plan).

méthode	supervision	annotation à l'inférence	scores (mIoU)	
			validation	test
<b>faiblement supervisé</b>				
BoxSup [4]	b.box		62.0%	64.2%
SEAM [24]	caption		64.5%	65.7%
<b>supervisé sans données additionnelles</b>				
ResNet-38 [25]	pixels		n.c.	82.5%
DeepLabv3+ [3]	pixels		81.63%	n.c.
<b>supervisé avec données additionnelles (COCO)</b>				
ResNet-38 [25]	pixels		80.84%	84.9%
DeepLabv3+ [3]	pixels		84.56%	89.0%

TABLEAU 4.2 – Scores de segmentation issus de la littérature sur les ensembles de validation/test de PASCAL VOC, pour des modèles entraînés sur l'ensemble d'apprentissage de PASCAL VOC (mIoU : moyenne sur les 21 classes).

plusieurs de ces critères permet alors de se rapprocher de la borne supérieure donnée par le mIoU, par exemple en combinant le critère sémantique au critère  $c_3$ . Le meilleur score moyen est alors obtenu avec le  $score_1$ , en utilisant à la fois les rectangles englobants et les labels. Cependant ajouter les labels n'apporte par un gain significatif contrairement aux rectangles englobants, ce qui atteste de l'intérêt supérieur d'annoter des rectangles englobants. Quant à lui, le  $score_2$  donne des résultats proches mais légèrement inférieurs. Enfin, une dernière solution consiste à remplir le rectangle englobant (à 90%) lorsque le score calculé est inférieur à un seuil (fixé à 0,5), ce qui permet d'atteindre la meilleure performance, très proche de la borne supérieure donnée par le mIoU.

Comme expérience supplémentaire "qualitative", nous avons utilisé cette chaîne pour segmenter *SpatialSense* [26], un jeu de données contenant des annotations de relations spatiales entre des objets représentés par leurs rectangles englobants (cf. Section 1.2.1.3). Ce jeu de données est composé de 11 569

images avec des objets de la vie quotidienne, des animaux, des personnes (avec plusieurs labels "homme", "femme", "fille"...), mais aussi des classes d'objets de type *stuff* (comme "ciel", "sol", "mur"...). Dans nos expérimentations, nous avons utilisé les images contenant certaines relations spatiales uniquement. La Figure 4.3 présente quelques exemples de segmentations obtenues sur ces données, en utilisant comme pré-segmentation un modèle HRNet entraîné sur le jeu de données ADE20K<sup>4</sup>. De tels résultats mettent en évidence l'intérêt de notre chaîne de traitement dans un cas réel, puisque *SpatialSense* est fourni sans aucune annotation de segmentation. Une telle segmentation peut être utile pour diverses tâches de vision par ordinateur, comme le calcul de descripteurs de position relative dans le cadre de cette thèse (cf. Chapitre 2).

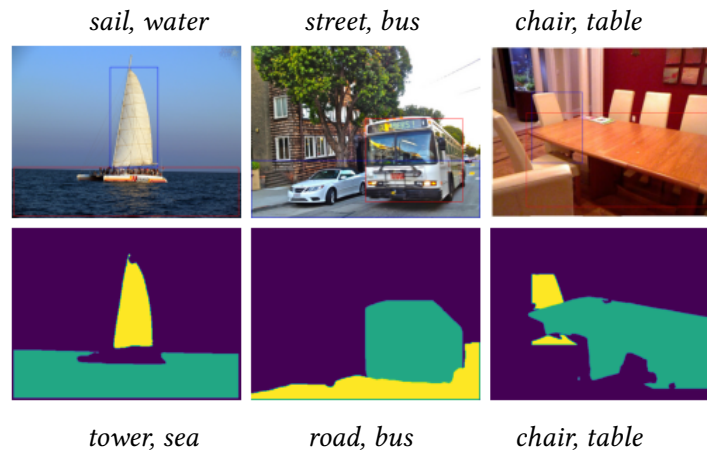


FIGURE 4.3 – Exemples de segmentations obtenues sur *SpatialSense*, avec pré-segmentation par HRNet entraîné sur ADE20K et sélection avec SegMyO.

(1<sup>e</sup> ligne) entrées : images + annotations (rectangle englobant + label);

(2<sup>e</sup> ligne) sorties : objets segmentés sélectionnés.

### 4.3 Conclusion

Nous avons introduit une chaîne de traitement pour extraire automatiquement des objets segmentés dans des images à l'aide de labels et/ou de rectangles englobants. Il s'appuie sur des critères simples pour sélectionner le meilleur segment parmi plusieurs propositions pour un objet donné, grâce à la similarité sémantique pour le label et à plusieurs mesures géométriques pour le rectangle englobant. Sa capacité à segmenter et sélectionner facilement et automatiquement tout objet dont on dispose de la classe et/ou du rectangle englobant en fait une solution pertinente pour segmenter un jeu de données sans nécessiter d'annotation dense ni d'entraînement spécifique. Par ailleurs, il s'agit également d'une amélioration intéressante pour les modèles de segmentation faiblement supervisés, que nous envisageons d'évaluer en l'intégrant dans un modèle dédié. Enfin, nous envisageons également d'évaluer plus précisément la complémentarité de nos critères et d'optimiser leur agrégation sur d'autres ensembles de données, où les différences sémantiques entre les classes pourraient être plus importantes.

Ces travaux ont été présentés dans un article de conférence internationale [5] et dans un atelier lors d'une conférence nationale [6]. Ils ont ensuite été étendus avec d'autres critères et une méthode de combinaison par apprentissage dans un autre article de conférence internationale [15]. Par ailleurs, une implémentation en *python* est disponible sur : <https://github.com/RobinDelearde/SegMyO>.

4. <https://github.com/CSAILVision/semantic-segmentation-pytorch>

## Chapitre 5

# Segmentation non supervisée basée sur la couleur

---

5.1	Introduction . . . . .	92
5.2	État de l’art . . . . .	92
5.2.1	Espaces de couleurs numériques polaires (HSX et HCX) . . . . .	93
5.2.2	Segmentation basée sur la couleur . . . . .	94
5.2.2.1	Segmentation d’images couleur . . . . .	95
5.2.2.2	Approches basées sur la couleur . . . . .	95
5.2.3	Classification de données directionnelles ou périodiques . . . . .	97
5.2.3.1	Comparaison de données périodiques . . . . .	98
5.2.3.2	Classification de données périodiques . . . . .	99
5.3	Segmentation dans les espaces HSX . . . . .	100
5.3.1	Approches proposées . . . . .	100
5.3.1.1	Classification basée sur la composante de teinte (H) . . . . .	100
5.3.1.2	Espace de mesure adapté pour les modèles HSX . . . . .	101
5.3.1.3	Classification dans l’espace de mesure adapté . . . . .	103
5.3.2	Expérimentations et résultats dans l’espace HSV . . . . .	104
5.3.2.1	Protocole expérimental . . . . .	105
5.3.2.2	Résultats et analyse . . . . .	106
5.4	Conclusion . . . . .	108

---

La segmentation d’images est une étape indispensable pour de nombreux traitements en compréhension de scènes, en permettant de définir précisément l’étendue des objets. Pour y parvenir, une des caractéristiques les plus discriminantes et évidentes est la couleur, à laquelle peut être associée la position des pixels. Le codage numérique des images couleur repose en général sur les trois canaux RGB, mais d’autres espaces de couleurs ont été développés, comme les modèles HSV/HSL, qui permettent de séparer une composante de chromaticité liée à la couleur perçue par la vision humaine et une composante de luminosité liée à la quantité de lumière, dans la lignée des atlas de couleurs comme ceux de Munsell ou Ostwald. Ces modèles nécessitent cependant un traitement particulier afin de respecter leur définition et d’éviter des erreurs d’interprétation. Or, les approches usuelles de segmentation basée sur la couleur ne prennent pas forcément en compte leurs spécificités, à savoir le lien entre les composantes et l’aspect angulaire de la composante H. Nous proposons ici une solution en utilisant un espace de mesure adapté, où les distances correspondent bien aux écarts de couleurs dans le modèle, ce qui permet d’utiliser une méthode de classification par *clustering* comme les *k*-moyennes. Nous avons évalué notre approche sur plusieurs séquences d’images et obtenu de bons résultats, ce qui nous a permis de l’utiliser dans une chaîne de traitements plus complexe. Nous montrons également que la pondération des

composantes dans cet espace permet de modifier l'importance donnée à la chromaticité ou à la couleur, ce qui peut être utile pour des traitements nécessitant une invariance à l'un ou à l'autre.

## 5.1 Introduction

La segmentation d'images est une tâche majeure en compréhension de scènes, à la base de nombreux traitements plus avancés. Elle permet d'extraire les différents objets présents dans l'image de façon plus précise que la simple détection, ce qui permet ensuite de mieux les décrire de façon automatique, que ce soit en termes de forme, de couleur ou de relations. Appliquée à un objet, elle peut aussi permettre de le décomposer en parties, pour des descriptions plus fines encore. En particulier, la couleur est une caractéristique déterminante pour distinguer des objets ou parties d'objets, et peut même suffire dans de nombreux cas. Elle est néanmoins sensible aux conditions d'acquisition de l'image, notamment en termes de luminosité, ce qui nécessite des traitements adaptés.

D'abord basées sur des images en niveaux de gris, lorsque les performances de calcul étaient très limitées, les premières solutions consistaient simplement en un seuillage sur ces niveaux de gris, avec des seuils fixés manuellement ou automatiquement à partir des distributions des pixels (histogrammes). La segmentation d'images couleur est apparue dans les années 1970 avec le début de la généralisation de l'utilisation de telles images. Celles-ci sont alors encodées selon 3 canaux RGB ou dans d'autres espaces de couleurs, comme les espaces HSV et HSL (notés HSX dans la suite) visant à distinguer chromaticité et luminosité, ou encore les espaces YUV ou YIQ pour la télévision. Dans le même temps, des approches de classification plus évoluées ont pu être développées grâce à l'amélioration des capacités de calcul, comme les approches par *clustering*. De nombreux travaux de recherche se sont intéressés à ces sujets et ont proposé d'utiliser ces méthodes pour segmenter les images couleur directement à partir de cette caractéristique, en se plaçant dans un espace de couleurs différent pour mieux la prendre en compte. Cependant, les approches courantes ne prennent pas bien en compte les spécificités de ces espaces, notamment des espaces HSX, qui nécessitent des solutions adaptées. De plus, ces approches semblent parfois avoir été oubliées, de nombreuses applications récentes (en santé ou en agriculture par exemple) continuant d'utiliser des solutions basiques à partir de seuillage [56, 43, 37, 38]. Tout ceci pousse à remettre en avant ces méthodes et à poursuivre les recherches dans ce domaine, en commençant par les évaluer avec les performances actuelles.

Dans l'objectif d'avoir une solution simple et efficace pour segmenter des images en fonction de la couleur, en distinguant chromaticité et luminosité, nous proposons d'utiliser une segmentation par *clustering* dans un espace de couleurs HSX (HSV dans nos expérimentations), en l'adaptant aux spécificités de cet espace. Cette approche peut alors servir à décomposer des objets en parties et être intégrée dans une chaîne de traitement plus large. Dans nos travaux, nous l'avons utilisée pour segmenter des personnes dans des images aériennes, pour une tâche de ré-identification (cf. Chapitre 6). Les contributions de ce chapitre sont les suivantes :

- analyse de l'utilisation des espaces de couleurs HSX pour la segmentation ;
- introduction de la notion d'espace de mesure adapté pour la comparaison des couleurs ;
- proposition et évaluation de plusieurs méthodes de segmentation basée sur la couleur dans l'espace HSV.

## 5.2 État de l'art

Nous donnons dans cette section un résumé de l'état de l'art en segmentation basée sur la couleur. Afin d'en comprendre les différentes problématiques, nous proposons d'abord une brève introduction des espaces de couleurs, notamment les espaces polaires (de type HSX). Nous abordons également la problématique de la classification de données périodiques, qui apparaît lorsque l'on travaille dans ce type d'espace, avec la composante de teinte (H).

### 5.2.1 Espaces de couleurs numériques polaires (HSX et HCX)

La perception de la couleur a fait l'objet de nombreuses études depuis plusieurs siècles, en sciences optiques puis en neurosciences d'une part, et pour des applications artistiques puis industrielles d'autre part. En particulier, la mesure de la perception humaine de la couleur est devenue un domaine scientifique à part entière, appelé colorimétrie. En science informatique, elle amène deux défis majeurs : retranscrire en grandeurs numériques les grandeurs physiques mesurées par les capteurs, et utiliser des représentations de données adaptées pour retranscrire la perception humaine. Dans cette optique, plusieurs espaces de couleurs numériques ont été développés, comme les espaces HSV et HSL, en s'appuyant sur les travaux antérieurs de catégorisation des couleurs sous forme d'atlas, comme ceux de Munsell ou Ostwald. Ceux-ci permettent de séparer la couleur en une composante chromatique (la teinte notamment) et une composante de luminosité, ce qui est particulièrement utile pour des tâches de reconnaissance où la luminosité peut varier. Nous proposons en Annexe A un bref historique du développement des espaces de couleurs, notamment les espaces polaires (de type HSX).

Dans le but de catégoriser les couleurs en fonction de leur perception par la vision humaine, de nombreuses approches, comme celles de Munsell ou Ostwald, ont produit des atlas 3D en décomposant ces couleurs selon trois grandeurs (cf. Annexe A) :

- une grandeur angulaire de chromaticité appelée teinte (H, de l'anglais "*hue*") correspondant à la position sur la roue des couleurs ;
- une grandeur correspondant à la vivacité de la couleur, soit absolue avec la chrominance (C, utilisée par Munsell), soit relative ou normalisée avec la saturation (S, utilisée par Ostwald) ;
- une grandeur correspondant à la luminosité, soit réelle avec l'intensité lumineuse (I) ou la luminance (L), soit ressentie avec la "valeur" (V) introduite dans le système de Munsell.

À noter que le terme de "brillance" (ou de *brightness* en anglais) n'est pas utilisé ici du fait de son ambiguïté, puisqu'il peut renvoyer à la luminosité réelle (L ou I), la luminosité perçue (V), ou encore la vivacité de la couleur (C ou S).

Il est important de souligner que la chrominance telle que définie par Munsell traduit directement la perception humaine de vivacité des couleurs, de façon absolue. De ce fait, certaines associations de teintes et valeurs présentent plus de variété de chrominance que d'autres, comme le montrent les Figures A.5 et A.2, ce qui donne des intervalles différents et une représentation 3D non uniforme. À l'inverse, le modèle d'Ostwald utilise la chrominance relative, ou saturation, en découpant l'espace en secteurs avec un écart identique suivant cette grandeur, ce qui donne une représentation uniforme (cf. Figure A.4).

Par la suite, ces grandeurs physiques ont été traduites en grandeurs numériques à partir des composantes RGB des pixels, selon les formulations données dans [19, 44, 16] notamment. Elles ont donné naissance à plusieurs modèles de couleurs selon les grandeurs utilisées, comme les modèles HSV, HSL et HSI [44, 16], que nous englobons sous l'appellation "HSX". Les définitions employées aujourd'hui sont données dans l'équation 5.1, avec des définitions différentes de la saturation selon la grandeur de luminosité utilisée (V, I ou L). Cependant, ces modélisations numériques ne retranscrivent pas exactement les systèmes de couleurs expérimentaux (comme l'atlas de Munsell), puisqu'elles sont basées sur des formules empiriques sur les composantes RGB, et non sur des mesures expérimentales. Elles dépendent donc des formules choisies d'une part, et de la définition des composantes RGB d'autre part, c'est-à-dire de la transcription des signaux mesurés par le capteur en valeurs numériques.

En particulier, une nouvelle définition de la chrominance est donnée ici, comme étant la différence entre la valeur maximale et la valeur minimale des composantes RGB de la couleur considérée. Les valeurs RGB couvrant toutes le même intervalle  $[0, 1]$  (ou  $[0, 255]$  pour des valeurs codées sur 8 bits), il en est de même pour cette différence. Pour une valeur  $V$  donnée, la chrominance maximale est égale à  $V$ , et elle est atteignable quelle que soit la valeur de teinte. Cette chrominance "numérique" est donc déjà une grandeur relative en comparaison avec la chrominance de Munsell, puisqu'elle est normalisée sur cet intervalle. Dans la définition du HSV, la saturation est alors définie comme la chrominance norma-



lisée par rapport à cette valeur, donc couvrant l'intervalle  $[0, 1]$  quelle que soit la valeur. Pour les autres grandeurs de luminosité (intensité ou luminance), les définitions de la saturation permettent d'avoir ce même comportement (iso-teinte et couvrant l'intervalle  $[0, 1]$ ). À noter que plusieurs modèles, regroupés sous l'appellation "HCX", utilisent cette chrominance numérique relative et non une chrominance absolue comme celle de Munsell. En réalité, ils sont donc équivalents aux modèles HSX correspondants, seule leur représentation étant différente.

D'autres modèles similaires ont été proposés ensuite, comme le GLHS (*Generalised Lightness, Hue and Saturation*) qui est une généralisation de ces différents espaces [22]. Par ailleurs, une analyse de ces espaces a été menée par Serra et Angulo, aboutissant à une définition plus adaptée au traitement d'images utilisant la norme  $\mathcal{L}_1$  [41, 2, 42, 1].

Avec  $(R, G, B) \in [0, 1]^3$ ,

$$\begin{aligned}
 V &= \max(R, G, B) \\
 I &= (R + G + B)/3 \\
 L &= (\max(R, G, B) + \min(R, G, B))/2 \\
 C &= \max(R, G, B) - \min(R, G, B) \\
 H &= 60^\circ \times \begin{cases} \text{non défini} & \text{si } C = 0 \\ (\frac{G-B}{C} + 0) \pmod{6} & \text{si } V = R \\ (\frac{B-R}{C} + 2) \pmod{6} & \text{si } V = G \\ (\frac{R-G}{C} + 4) \pmod{6} & \text{si } V = B \end{cases} \\
 S_V &= \begin{cases} 0 & \text{si } V = 0 \\ \frac{C}{V} = 1 - \frac{\min(R, G, B)}{V} & \text{sinon} \end{cases} \\
 S_I &= \begin{cases} 0 & \text{si } I = 0 \\ 1 - \frac{\min(R, G, B)}{I} & \text{sinon} \end{cases} \\
 S_L &= \begin{cases} 0 & \text{si } L = 0 \text{ ou } L = 1 \\ \frac{C}{1-|2L-1|} = \frac{\max(R, G, B) - \min(R, G, B)}{1 - |\max(R, G, B) + \min(R, G, B) - 1|} & \text{sinon} \end{cases}
 \end{aligned} \tag{5.1}$$

Tous ces modèles sont très utiles pour décrire la couleur de façon instinctive et avec des composantes indépendantes, en dissociant la chromaticité de la luminosité. Mais en réalité, ces définitions à partir du modèle RGB ne permettent pas de couvrir l'ensemble des couleurs visibles par l'œil humain. D'autres modèles plus proches de la perception humaine peuvent alors être utilisés, comme le  $L^*a^*b^*$  ou le  $L^*u^*v^*$  (aussi connus sous les noms de CIELAB et CIELUV) qui ont été définis par la CIE en 1976 à partir de l'espace XYZ et des travaux de Hunter et MacAdam notamment [13, 25]. Ces espaces sont en fait équivalents au HCL, en redonnant à la chrominance une définition plus proche de sa signification physique. La teinte et la chrominance sont alors les coordonnées polaires de ces espaces (angle et norme), tandis que  $(a^*, b^*)$  ou  $(u^*, v^*)$  sont les coordonnées cartésiennes. Les unes peuvent donc être obtenues à partir des autres, ce qui est par exemple exploité dans [8]. Enfin, d'autres transformations non liées à un espace de couleurs particulier peuvent être utilisées afin de réduire la corrélation entre les composantes, comme la transformation de Karhunen-Loeve (KLT) étudiée notamment dans [31, 21].

### 5.2.2 Segmentation basée sur la couleur

La segmentation d'images est une tâche majeure en vision par ordinateur et a été largement traitée depuis de nombreuses années, comme l'importante littérature sur le sujet peut en témoigner, de nombreux états de l'art étant disponibles également [33, 24, 10, 52]. Plusieurs tâches sont aussi très proches de ce sujet, comme la proposition d'objets, la détection de "blobs", régions ou superpixels, la détection

de contours, ou encore la segmentation objet-fond. Par ailleurs, la segmentation basée sur la couleur est elle-même un sous-domaine important, qui est intimement lié à celui des espaces de couleurs. Ainsi, plusieurs états de l'art sont dédiés à ces aspects, comme ceux de [5, 12], tandis que [6] étudie plus spécifiquement la segmentation dans l'espace HSL. Nous résumons ici les principales approches puis détaillons celles basées sur la couleur.

### 5.2.2.1 Segmentation d'images couleur

Les principales approches de segmentation sont les suivantes :

- la détection de contours, en utilisant des filtres adaptés ou la ligne de partage des eaux (*watershed*) par exemple,
- la classification non supervisée sur des attributs des pixels (couleur, position) ou des régions (couleur, position, texture), avec de nombreuses méthodes :
  - le simple seuillage, selon des attributs de niveaux des couleurs typiquement, en fixant les seuils manuellement ou automatiquement sur les histogrammes 1D ;
  - la croissance ou l'érosion de régions, en agrandissant ou réduisant des régions initiales selon leur similarité de manière incrémentale, afin d'obtenir des régions uniformes ;
  - le partitionnement (*clustering*), en cherchant à maximiser la distance entre régions et minimiser la distance interne des régions (ou des classes), de façon itérative typiquement, avec des techniques comme les *k*-moyennes, *mean shift*, le regroupement hiérarchique, le partitionnement de graphes ou encore les CNN ;
- l'apprentissage supervisé à partir d'annotations denses (classe par pixel), avec différentes méthodes plus ou moins supervisées :
  - par des arbres de décision, des SVM, des réseaux de neurones (MLP), des champs aléatoires conditionnels (CRF)... à partir des mêmes attributs que précédemment ;
  - par réseaux de neurones convolutionnels (CNN) directement sur les images, ce qui nécessite de grandes bases de données annotées, avec de multiples modèles comme *U-Net*, *Mask R-CNN*, *ResNet*, *DeepLab*...

La classification non supervisée a l'avantage de pouvoir adapter les classes de sortie aux données d'entrée, tandis que l'apprentissage supervisé permet d'apprendre en même temps à reconnaître le type d'objets, ajoutant de la sémantique et se rapprochant d'une tâche de reconnaissance. Ainsi, les modèles de segmentation par CNN sont souvent basés sur un réseau initial ("*backbone*") de classification d'objets, qui permet d'extraire automatiquement les caractéristiques pertinentes à partir de l'image brute. Les modèles d'apprentissage supervisé antérieurs n'ont en revanche pas permis d'atteindre d'aussi bonnes performances pour cette tâche de segmentation sémantique.

Par ailleurs, il est important de noter que la segmentation peut ne pas être une décision tranchée avec une seule classe par pixel et des frontières nettes entre les régions, mais peut avoir un comportement flou, avec des degrés d'appartenance non nuls pour plusieurs classes. C'est pourquoi des solutions basées sur la logique floue ont été proposées pour différentes méthodes, comme pour le *clustering* [14] ou la détection de contours [7].

### 5.2.2.2 Approches basées sur la couleur

Pour la segmentation basée sur la couleur, quatre approches sont principalement utilisées :

- la classification par seuillage, manuel ou automatique,
- la classification par croissance ou érosion de régions,
- la classification par *clustering* (ou partitionnement),
- la classification par apprentissage supervisé (avec ou sans sémantique).

Nous nous focalisons ici sur les approches non supervisées, a priori plus efficaces et plus adaptables aux données.



### Approches par seuillage et croissance de régions

L'approche la plus simple consiste à appliquer un seuillage sur les valeurs des différentes composantes de l'espace de couleurs retenu en analysant leurs distributions, soit de manière manuelle, soit de manière automatique. Plusieurs méthodes ont été proposées pour automatiser cela, comme celles d'Ohlander [30, 29] ou d'Otsu [32]. La première consiste à rechercher les pics des histogrammes de manière itérative, tandis que la seconde cherche à maximiser la séparabilité des classes en maximisant la variance inter-classes, en évaluant tous les seuils possibles. Introduites pour des images en niveaux de gris, ces techniques ont ensuite été étendues à plusieurs canaux en fixant différents seuils, dans différents espaces de couleurs. Ainsi, Tominaga définit les seuils de manière incrémentale en sélectionnant le pic le plus important parmi les trois histogrammes, à la manière de la méthode d'Ohlander. Il utilise cette solution dans les espaces HCV [48] et  $L^*a^*b^*$  [50]. Celenk propose quant à lui de traiter les composantes de manière indépendante, puis d'utiliser le discriminant de Fisher pour fusionner les sorties dans une seule dimension [8]. Il utilise cette solution dans l'espace HCL, dont les composantes sont obtenues à partir de l'espace  $L^*a^*b^*$ . Le seuillage, que ce soit automatique ou manuel, est encore utilisé pour de nombreuses applications, par exemple en agriculture [56], pour la navigation autonome [43, 37] ou la santé [38].

La croissance de régions consiste quant à elle à regrouper les pixels ou régions proches s'ils sont similaires, en partant d'un ou plusieurs points initiaux ("racines") et jusqu'à ce qu'il n'y ait plus de régions similaires ou que tous les pixels aient été affectés à une région. Cependant cette méthode dépend beaucoup de l'initialisation et demande beaucoup de ressources de calcul. Une autre méthode appelée "*split and merge*" consiste à diviser l'image en secteurs de plus en plus petits tant qu'ils ne sont pas homogènes, puis à fusionner les secteurs similaires, ce qui peut être modélisé par une structure d'arbre quaternaire ou de graphe. L'efficacité de cette méthode est dépendante de la forme des objets : plus il y a de variations dans l'image et plus elle demandera de divisions.

### Approches par *clustering*

Enfin, la classification par *clustering* consiste à trier les pixels de façon à maximiser la distance entre classes et minimiser la distance à l'intérieur des classes, en traitant tous les pixels dans une seule passe, souvent avec plusieurs itérations. Une méthode classique est celle des  $k$ -moyennes [26], qui définit la classe d'un pixel en fonction de  $k$  centroïdes, en lui donnant comme classe celle du centroïde le plus proche. Les centroïdes peuvent être initialisés de manière aléatoire, en prenant  $k$  éléments parmi les données d'entrée typiquement, ou de manière arbitraire. Ils sont ensuite mis à jour en les remplaçant par la moyenne de chaque classe, qui peut ne pas appartenir aux données existantes. Ce processus est alors répété jusqu'à ce que les classes ne changent plus, ou que la distance totale aux centroïdes soit en-dessous d'un seuil. Cette méthode dépend donc de trois paramètres : le nombre de classes  $k$ , l'initialisation et ce seuil de convergence. Le paramètre  $k$  optimal peut être obtenu automatiquement par la méthode du coude, tandis que les autres paramètres doivent passer par une analyse experte. D'autres méthodes de *clustering* sont également couramment utilisées, comme les algorithmes *mean-shift*, *DB-SCAN*, le regroupement hiérarchique ou le partitionnement de graphes.

Du fait des faibles performances de calcul aux débuts du traitement d'images, les approches par *clustering* ne pouvaient pas être appliquées directement dans l'espace de couleurs 3D (RGB ou autre), et l'état de l'art sur les images couleur reposait sur les méthodes par seuillage ou détection de contours. De ce fait, la segmentation par *clustering* a d'abord été appliquée aux images en niveau de gris, avec des performances limitées, ou en traitant indépendamment les composantes de l'espace de couleurs et en combinant les sorties, comme dans le cas du seuillage. Sur les images en niveau de gris, une idée commune est alors d'associer aux niveaux de gris des contraintes spatiales en utilisant les coordonnées des pixels, par exemple avec un champ aléatoire de Gibbs, qui peut être vu comme une généralisation des  $k$ -moyennes [35, 34]. Sur les images couleur, il est alors intéressant de travailler dans des espaces

de couleurs adaptés à la tâche, par exemple en utilisant la composante de teinte des espaces HSX/HCX. Cependant, cette composante seule n'est pas suffisante, comme l'ont confirmé expérimentalement les travaux de [54].

Avec l'augmentation des performances de calcul, le *clustering* a ensuite pu être appliqué aux images couleur directement dans l'espace 3D. Ainsi, [36] propose une méthode alternative aux  $k$ -moyennes basée sur des opérations de morphologie mathématique directement dans l'espace de couleurs 3D. Plusieurs espaces de couleurs sont évalués (RGB, XYZ, YIQ, UVW, I1-I2-I3) et la méthode est comparée à un  $k$ -moyennes dans l'espace RGB 3D, voire dans l'espace 5D incluant les coordonnées  $(x, y)$  des pixels. Afin de mieux prendre en compte la couleur, [54] se place quant à lui dans l'espace HSI et utilise deux  $k$ -moyennes : l'un sur la composante H et l'autre sur les composantes S et I, montrant l'importance de chacune des composantes, et suggérant d'utiliser la méthode de Pappas [35, 34] pour prendre en compte l'aspect spatial.

Plus tard, Angulo et Serra se sont également intéressés aux représentations polaires et ont proposé un espace de mesure plus adapté pour le traitement d'images, notamment pour la segmentation [2, 1]. Mais la plupart des approches ultérieures se sont plutôt tournées vers l'espace  $L^*a^*b^*$ , plus représentatif de la vision humaine. Ainsi, [4] utilise un algorithme de colonies de fourmis pour réaliser le *clustering* à l'aide des distances définies dans cet espace. D'autres approches proposent de combiner les attributs de couleur  $L^*a^*b^*$  avec d'autres informations, comme des attributs de texture [55] ou spatiaux [17]. Afin de traiter l'espace non euclidien engendré, [55] suggère d'utiliser l'algorithme des  $k$ -médoïdes plutôt que celui des  $k$ -moyennes, associé à la distance de Mallows. [17] propose quant à lui une amélioration de l'algorithme *mean-shift* dans l'espace de couleurs HSI, qu'il compare à l'algorithme des  $k$ -moyennes sur l'espace  $L^*a^*b^*$ , en y ajoutant éventuellement les coordonnées  $(x, y)$  des pixels.

Aujourd'hui, la segmentation sémantique et les approches par réseaux de neurones profonds ont largement pris le devant de la scène, y compris pour des solutions non supervisées [18, 15, 20]. Cela dit, la segmentation par *clustering* dans des espaces de couleurs trouve encore de nombreuses applications, par exemple en agriculture [23, 47].

### À propos du choix de l'espace de couleurs et de l'usage de la teinte

On observe donc que le choix du meilleur espace de couleurs pour la segmentation est un problème de longue date. De nombreuses solutions utilisent les espaces HSX ou  $L^*a^*b^*$  pour leur capacité à mieux représenter la chromaticité, en la séparant de la luminosité. La composante de teinte des espaces HSX est particulièrement utile pour caractériser la couleur d'un objet, mais ces espaces ne sont pas faciles à appréhender du fait de leurs spécificités [10, 5, 51, 12]. Ces spécificités sont les suivantes :

1. ce sont des espaces non euclidiens, ne permettant donc pas de mesurer des différences en calculant directement des distances euclidiennes ;
2. les composantes sont liées entre elles par des contraintes dues à leur définition, générant des "singularités" ;
3. la teinte est une composante angulaire, donc périodique, et doit donc être traitée comme telle.

Ainsi, bon nombre de travaux ne tiennent pas compte de la particularité de ces espaces, et en particulier de l'aspect angulaire de la teinte, ce qui peut générer de grandes erreurs d'appréciation des résultats obtenus. La Section 5.2.3 approfondit ce sujet et propose des pistes pour traiter correctement cette grandeur. À noter cependant que plusieurs approches le font déjà, comme celles de [8, 51, 54, 1, 4], ce qui mérite d'être souligné.

### 5.2.3 Classification de données directionnelles ou périodiques

La classification de données périodiques est un problème très peu abordé par la littérature, bien que de nombreuses données de ce type existent. On peut distinguer deux principaux types de données périodiques : celles contenant des grandeurs périodiques et les signaux périodiques. Les premières

font référence à des données pour lesquelles au moins une dimension est une grandeur (scalaire) périodique, comme un angle, la phase d'un signal temporel, la teinte dans l'espace de couleurs HSV, ou encore les différentes grandeurs modélisées par les "capsules" dans les réseaux du même nom [39]. On emploie également le terme de "direction" pour désigner ces grandeurs lorsqu'elles caractérisent une donnée multi-dimensionnelle (vecteur), par opposition à la norme de celle-ci, ce qui en fait des données dites directionnelles. Le deuxième type de données périodiques concerne les signaux périodiques, qu'ils soient périodiques parce qu'ils sont définis en fonction d'un seul paramètre périodique (un angle par exemple), comme un histogramme d'angles (cf. chapitre 2), ou parce qu'ils se répètent selon un paramètre non périodique (le temps par exemple), comme l'amplitude d'un signal lumineux. Par extension, on peut également considérer des signaux qui évoluent suivant un paramètre périodique et d'autres paramètres non périodiques sans que ces signaux soient forcément périodiques, comme la consommation de gaz au cours des mois de l'année dans l'évaluation de [53].

Pour comparer et classifier de telles données, il est nécessaire de prendre en compte leur périodicité. Plusieurs solutions sont alors possibles suivant le type de données.

### 5.2.3.1 Comparaison de données périodiques

#### Comparaison de grandeurs périodiques

Pour comparer directement les valeurs de grandeurs périodiques (ou directionnelles), il suffit de calculer leur différence angulaire, en adaptant la période. Ce calcul est facilement effectué par n'importe quel outil de calcul, en définissant le type de la donnée pour se placer dans un espace adapté, plutôt qu'en redéfinissant les règles de calcul comme dans [27, 53].

Une autre solution proposée par [53] consiste à transformer les valeurs angulaires par des points du cercle unité, grâce aux fonctions trigonométriques, de façon à pouvoir calculer des distances classiques sur ces points dans l'espace 2D, comme la distance euclidienne dans l'équation 5.2. On passe donc de grandeurs scalaires périodiques à des grandeurs 2D non périodiques. Cependant, cette solution a le défaut d'introduire des distorsions selon [53], notamment avec la distance euclidienne (norme  $\mathcal{L}_2$ ). Une première amélioration suggérée dans [2] serait d'utiliser la distance de Manhattan (norme  $\mathcal{L}_1$ ), comme dans l'équation 5.3, mais elle n'a pas été évaluée ici.

$$\text{Avec } (\theta_1, \theta_2) \in \mathbb{R}^2 \text{ et } \begin{cases} x_i = \cos(\theta_i) \\ y_i = \sin(\theta_i) \end{cases},$$

$$\begin{aligned} d_{trigo, \mathcal{L}_2}(\theta_1, \theta_2) &= \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \\ &= \sqrt{(\cos(\theta_1) - \cos(\theta_2))^2 + (\sin(\theta_1) - \sin(\theta_2))^2} \end{aligned} \quad (5.2)$$

$$\begin{aligned} d_{trigo, \mathcal{L}_1}(\theta_1, \theta_2) &= |x_1 - x_2| + |y_1 - y_2| \\ &= |\cos(\theta_1) - \cos(\theta_2)| + |\sin(\theta_1) - \sin(\theta_2)| \end{aligned} \quad (5.3)$$

Une mesure de la similarité (comprise entre 0 et 1) entre deux valeurs périodiques peut aussi être obtenue en prenant la valeur absolue du cosinus de leur demi-différence. La similarité sera alors maximale pour des valeurs proches ou décalées de  $2\pi$ , et minimale pour des valeurs décalées de  $\pi$ . Ceci permet aussi d'en déduire une distance sur de telles données, qui sera comprise entre 0 et 1 (cf. équation 5.4). Cependant cette fonction n'est pas linéaire, contrairement aux solutions précédentes, ce qui peut être problématique dans certains cas.

$$\text{Avec } (\theta_1, \theta_2) \in \mathbb{R}^2, \quad d_{\cos}(\theta_1, \theta_2) = 1 - \left| \cos\left(\frac{\theta_1 - \theta_2}{2}\right) \right| \quad (5.4)$$

### Comparaison de données directionnelles multidimensionnelles

Pour comparer des données directionnelles multidimensionnelles, où la direction n'est pas donnée par une seule valeur, il est nécessaire d'utiliser d'autres méthodes, ce qui a été étudié en statistiques notamment [27]. L'approche la plus populaire est de mesurer le produit scalaire (cf. équation 5.5) ou la similarité cosinus (cf. équation 5.6). Ceci permet d'exploiter facilement la direction des vecteurs, y compris sur des vecteurs de grandes dimensions, le produit scalaire pouvant être calculé analytiquement à partir des coordonnées des vecteurs. Par ailleurs, il a été vérifié expérimentalement que de telles mesures sont également plus adaptées sur des données textuelles représentées par des vecteurs de concepts de grande dimension [45], bien que le sens physique de la direction de ces vecteurs soit plus abstrait. Enfin, on peut noter également l'existence d'une mesure de distance dédiés aux histogrammes sur des données périodiques, pouvant par exemple être utilisée sur des histogrammes de valeurs de teinte [9].

Avec  $\vec{u}_1 \in \mathbb{R}^N$  et  $\vec{u}_2 \in \mathbb{R}^N$ ,

$$\vec{u}_1 \cdot \vec{u}_2 = \|\vec{u}_1\| \cdot \|\vec{u}_2\| \cdot \cos(\vec{u}_1, \vec{u}_2) = \sum_{j=1}^N u_{1,j} \cdot u_{2,j} \quad (5.5)$$

$$s_{\cos}(\vec{u}_1, \vec{u}_2) = |\cos(\vec{u}_1, \vec{u}_2)| = \frac{|\vec{u}_1 \cdot \vec{u}_2|}{\|\vec{u}_1\| \cdot \|\vec{u}_2\|} \quad (5.6)$$

#### 5.2.3.2 Classification de données périodiques

Bien que de nombreuses données de ce type existent, la classification de données périodiques est un problème très peu abordé par la littérature, que ce soit pour des approches non supervisées de type *clustering* ou supervisées avec des réseaux de neurones par exemple. Quelques approches existent néanmoins pour différentes tâches et différents types de données périodiques (simples grandeurs, direction de données multi-dimensionnelles ou signaux périodiques).

Concernant la classification de grandeurs périodiques, les travaux de [53] ont cherché à adapter la méthode des *k*-moyennes [26] à de telles données, visant entre autres le cas d'usage de la classification de la teinte dans l'espace de couleurs HSV. Deux solutions sont identifiées : un encodage trigonométrique en deux dimensions (*cosinus* et *sinus*) pour utiliser une distance classique (cf. équation 5.2), ou une simple différence angulaire (dont le calcul est ré-inventé dans l'article). Compte-tenu des potentiels défauts introduits par la première solution, [53] recommande d'utiliser la seconde, ce qui est vérifié sur un cas de test de données temporelles quasi périodiques (consommation de gaz au cours d'une année).

Pour les données multi-dimensionnelles, quelques approches de *clustering* ont aussi été développées pour des données textuelles représentées par des vecteurs de concepts de grande dimension, soit en utilisant des mesures de similarité comme la similarité cosinus [45], soit en développant des approches nouvelles comme les *k*-moyennes sphériques [11] ou celle de [3] basée sur des distributions de von Mises-Fischer sur l'hypersphère unité. Cependant ces dernières n'ont jamais été appliquées à des données image à notre connaissance. Par ailleurs, une autre approche sur de telles données est celle des réseaux à capsules, où la comparaison entre deux capsules est effectuée grâce à un produit scalaire [39].

Enfin, sur des signaux périodiques, une solution est d'utiliser des CNN avec un *padding* circulaire, comme nous l'avons fait avec l'histogramme (ou le bandeau) de forces (cf. Section 2.2.3). Ceci consiste à ajouter des zéros ou des données interpolées sur les bords de l'image lors du passage des filtres de convolution, afin d'obtenir une réponse à ces filtres et ainsi avoir une chance de détecter les motifs recherchés sur les bords également. Cette opération est adaptée à tout type de données périodiques, comme les signaux 1D (avec des convolutions 1D) ou les images panoramiques (360°) [40].

## 5.3 Segmentation dans les espaces HSX

### 5.3.1 Approches proposées

Dans l'objectif d'avoir une solution simple et efficace pour segmenter des images en fonction de la couleur, nous proposons d'utiliser une segmentation par partitionnement ("*clustering*") des pixels dans un espace de couleurs HSX. Ce choix fait suite à l'étude de la solution Magellium pour la décomposition en parties (cf. Chapitre 6), qui consistait en un échantillonnage des couleurs basé sur la teinte du modèle HSV pour les "vraies" couleurs, s'inspirant pour cela du système de Munsell, et sur la composante K du modèle CMYK pour le noir et le blanc, avec une étape de fusion des deux sorties. En effet, à première vue, l'espace HSV permet de traiter la couleur en dissociant une composante propre à la nature des objets (la teinte H) de composantes davantage liées aux conditions de vue (la valeur V, liée à la luminosité, et de façon plus limitée la saturation S). Utiliser la composante H seule pourrait donc permettre d'avoir une solution robuste aux conditions de vue, ou tout du moins aux variations de S et V. Cependant, la composante H seule ne permet pas de traiter les "couleurs" grises, noires et blanches, et en réalité les composantes S et V peuvent aussi dépendre de la nature des objets. L'utilisation de la composante K du modèle CMYK peut permettre de traiter le premier problème, mais par le second. De plus, le traitement par échantillonnage ne permet pas de s'adapter aux données, en fixant des seuils arbitraires susceptibles de séparer des pixels de couleurs très proches appartenant à un même objet.

Considérant que les modèles HSX sont malgré tout intéressants pour analyser les couleurs des objets, ainsi que pour apporter de l'invariance aux conditions de vue, et plus pratiques que le modèle RGB par exemple, nous proposons ici de les utiliser mais avec un traitement différent considérant les trois composantes, sans utiliser d'autre modèle. Pour cela nous suggérons de donner une importance supérieure à la composante H, et de traiter les données dans un espace de représentation adapté, prenant en compte l'aspect périodique de cette composante notamment. Le *clustering* permet alors d'avoir une solution non supervisée qui s'adapte aux données, plutôt que de définir des seuils arbitrairement et manuellement pour chaque image. Nous proposons d'utiliser la méthode des *k*-moyennes, qui donne de bonnes performances en général et est simple d'utilisation.

Dans la suite, nous proposons une approche de classification de pixels par *clustering* adaptée aux modèles HSX, en évaluant d'abord une solution basée sur la teinte seule, puis une solution séquentielle, et enfin dans un espace de mesure adapté pour le HSV.

#### 5.3.1.1 Classification basée sur la composante de teinte (H)

De nombreuses approches de segmentation basée sur la couleur utilisent la composante de teinte (H), qui est présente dans les modèles HSX ou HCX. Cette grandeur est définie comme un angle en se basant sur le concept de roue des couleurs, ce qui permet de décrire les couleurs principales en modélisant l'aspect cyclique et continu de la perception des couleurs, ainsi que la notion de couleurs opposées (cf. Figure A.1). Les autres composantes permettent alors de décrire toutes ses variantes en termes de luminosité (du noir au blanc en passant par la couleur) ou de valeur (du noir à la couleur vive), et de saturation (du gris à la couleur vive), comme sur la coupe de la Figure A.2. La grandeur de teinte a alors un important pouvoir de discrimination des couleurs et permet de réaliser des traitements de segmentation sans tenir compte des variations de luminosité, ce qui a un intérêt majeur pour de nombreuses applications [49].

Ainsi, de nombreuses approches basiques utilisent cette grandeur seule pour mesurer la proximité entre deux couleurs. Cependant, la plupart traitent la teinte comme une grandeur classique, alors qu'il s'agit d'un angle. Ceci génère une différence importante entre les couleurs situées de part et d'autre de la valeur 0 (le rouge), alors qu'en réalité elles sont proches sur le cercle. Un traitement particulier adapté aux données périodiques est donc nécessaire pour cette grandeur.

La solution la plus simple pour classifier des données angulaires comme la teinte des pixels consiste

à diviser le cercle en secteurs angulaires de même taille, ce qui revient à réaliser un échantillonnage des valeurs. C'est la méthode de description des teintes utilisée par le modèle de Munsell, avec un nombre de teintes fixé à 8 par exemple dans la Figure A.2, ou 10 dans la Figure A.5, ainsi que dans la chaîne de traitement Magellium originale pour la ré-identification. Cependant ce seuillage fixe ne permet pas de s'adapter aux données, ce qui peut être problématique pour une tâche de classification. En effet, deux valeurs de H très proches peuvent se retrouver dans des secteurs différents, tandis que des valeurs beaucoup plus éloignées sont dans le même secteur. De plus, s'il y a peu de teintes différentes dans les données, plusieurs secteurs peuvent être inutiles, tandis qu'il serait plus utile d'avoir un découpage plus fin là où il y a des données. Une solution s'adaptant aux données comme le *clustering* est donc plus intéressante pour cela.

Très peu de solutions existent pour le *clustering* de grandeurs périodiques, et aucune n'a été proposée pour les données image à notre connaissance, si ce n'est celle de [53] pour l'algorithme *k*-moyennes, mais qui ne présente pas de résultats sur de telles données (cf. Section 5.2.3). Deux solutions sont en fait proposées : un encodage trigonométrique en 2 dimensions (*cosinus* et *sinus*) ou une modification du calcul de la distance et des centroïdes. Du fait des distorsions possibles avec la première méthode, la seconde est recommandée. Les résultats expérimentaux de [53] sur des données non image montrent que celle-ci est efficace et évite les défauts de la première. Dans un premier temps nous proposons donc d'utiliser cette méthode pour la classification de la teinte (H) dans des images. L'implémentation de l'algorithme des *k*-moyennes modifié se fait facilement, et notre version en *python* sera diffusée dès que possible. Pour le calcul des distances et des centroïdes, nous suggérons de passer en coordonnées complexes pour simplifier le traitement, comme défini dans l'équation 5.7.

Avec  $(\theta_1, \dots, \theta_N) \in \mathbb{R}^N$ ,

$$d_{uni}(\theta_i, \theta_j) = |\theta_i - \theta_j| \pmod{2\pi}$$

$$\theta_{centroid} = \text{angle}\left(\sum_{i=1}^N e^{j\theta_i}\right) \quad (5.7)$$

Une approche hybride permet alors de combiner une grandeur périodique avec des grandeurs "classiques", comme les différentes composantes du modèle HSV, ce que nous proposons d'utiliser. Une solution naïve consiste à additionner les distances obtenues pour chaque coordonnée, en normalisant celles-ci ou en les pondérant éventuellement pour donner un poids plus ou moins important à certaines composantes. Les deux solutions pour le H donnent alors deux solutions pour la combinaison : soit avec les 4 dimensions (*cos H*, *sin H*, *S*, *V*) dues aux deux dimensions du H, soit simplement avec les 3 dimensions (*H*, *S*, *V*) en utilisant la différence angulaire pour le H.

Cependant il n'est pas toujours justifié de traiter la teinte comme une caractéristique utile pour la classification. En effet, celle-ci n'a pas beaucoup de sens pour les couleurs dont la saturation est faible. En particulier, utiliser la teinte seule ne permet pas de traiter correctement les nuances "achromatiques" (noir, gris, blanc), où la saturation est faible. De même, la saturation n'a pas de sens pour les couleurs dont la valeur est faible, donc utiliser la saturation telle quelle n'est pas toujours pertinent non plus. Afin d'utiliser correctement ces grandeurs, il est nécessaire de considérer leur définition pour bien prendre en compte les liens entre elles. Ces définitions peuvent se traduire par une modélisation géométrique, comme détaillé ensuite, donnant une solution plus adaptée pour la combinaison.

### 5.3.1.2 Espace de mesure adapté pour les modèles HSX

La représentation choisie pour l'espace de couleurs et la distance associée ont leur importance lorsque l'on effectue des comparaisons dans cet espace, puisqu'elle traduisent plus ou moins bien les grandeurs physiques en grandeurs numériques et les écarts dans le modèle physique en différences numériques, ce qui est particulièrement important pour les problèmes de classification [31, 50]. Par exemple, le modèle RGB peut être représenté par un espace cubique, où chaque composante est utilisée



comme une coordonnée cartésienne (cf. Figure 5.1). Une couleur est alors une combinaison linéaire des trois composantes, traduisant le principe d'additivité des couleurs à partir des trois couleurs primaires, et une différence dans cet espace traduit un écart de couleur équivalent suivant la définition du RGB.

De la même façon, la plupart des approches traitent les composantes des modèles de couleurs primaires (HSV par exemple) comme s'il s'agissait de coordonnées cartésiennes classiques, donc dans un espace cubique. Or, par définition ces modèles ont une coordonnée angulaire et présentent des liens particuliers entre les composantes, ce qui ne permet pas d'utiliser un tel espace pour la mesure de distance entre couleurs. D'autres représentations sont alors nécessaires pour prendre en compte ces spécificités. Nous introduisons ici la notion d'espace de mesure adapté pour un modèle de couleurs, qui est une représentation géométrique traduisant directement les différences de couleurs par les distances entre les points correspondants dans cet espace. Cela renvoie également à la notion d'*Uniform Color Scale* (UCS) [46], dont le but est de traduire mathématiquement et directement une différence de perception en distance dans l'espace de couleurs, par une distance euclidienne typiquement.

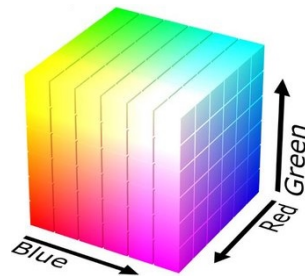


FIGURE 5.1 – Représentation cubique de l'espace de couleurs RGB.

L'équation 5.1 rappelle la définition mathématique des composantes HSV, HSL et HSI à partir du RGB. On peut constater que les valeurs ne sont pas indépendantes, ce qui impose des contraintes si l'on souhaite représenter l'espace des valeurs possibles de façon réaliste, en utilisant les 3 composantes comme dimensions (avec une dimension angulaire pour le H). En fait, la définition mathématique de ces grandeurs conduit à des espaces 3D avec des géométries bien spécifiques. Le fait d'avoir une composante angulaire impose d'avoir une coordonnée circulaire, donc un espace comme un ellipsoïde, un cylindre ou un cône. Le lien entre S et V conduit alors à restreindre cet espace à un cône ou un demi-ellipsoïde pour le HSV, tandis que le lien entre S et L ou I conduit à un double-cône (ou "bi-cône") ou un ellipsoïde pour le HSL et le HSI. La coordonnée orthogonale à l'axe traduit alors la variation de saturation, parcourant le nouvel espace de 0 à 1, tandis que la coordonnée selon l'axe traduit la variation de la valeur ou de la luminance, comme représenté sur la Figure 5.2.

Ces représentations sont souvent utilisées pour la visualisation de ces espaces mais pas pour le calcul, où la majorité des approches utilisent en fait un espace cubique en utilisant directement des métriques classiques sur les coordonnées cartésiennes. Ceci est erroné pour plusieurs raisons : premièrement, H est périodique, donc il doit être traité comme une grandeur périodique, en le représentant par un angle typiquement ; deuxièmement, H n'a pas de sens lorsque S est nul, donc mesurer des différences sur H n'a pas de sens dans ce cas, bien qu'une valeur lui soit tout de même assignée dans les images numériques. Par ailleurs, le cylindre est aussi couramment utilisé pour la visualisation des espaces HSX, mais cette représentation n'est pas correcte pour traduire les distances entre couleurs et l'importance de chaque couleur, en déformant la perception encodée par le modèle [41]. En effet, elle donne plus de volume aux couleurs avec un S ou V (ou I ou L) faible, ce qui donne plus d'importance à ces couleurs, pourtant moins nombreuses dans notre perception et dans l'espace RGB. De plus, elle autorise plusieurs points pour le noir (et pour le blanc avec le HSL et le HSI), avec des valeurs différentes pour les composantes H et S, ce qui n'est pas conforme à la définition et génère des distances entre ces points alors qu'ils encodent la même couleur. La mesure de distances nécessite donc de se placer dans un espace adapté, comme l'espace conique pour le HSV, ou bi-conique pour le HSL et le HSI.

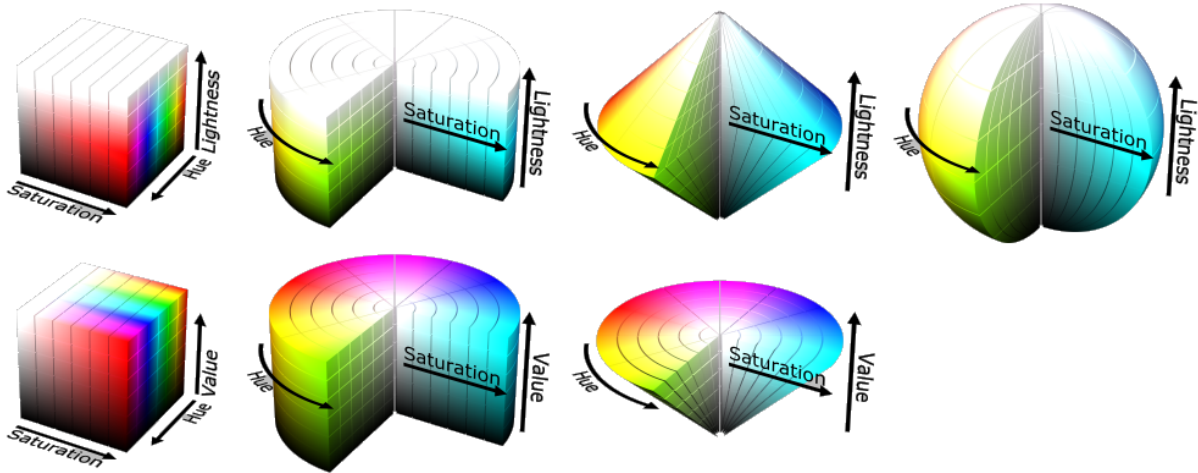


FIGURE 5.2 – Exemples de représentations des espaces de couleurs avec les modèles HSL (en haut) et HSV (en bas). Plusieurs représentations visuelles sont possibles, en utilisant une traduction différente des composantes colorimétriques en coordonnées de l'espace, mais toutes ne traduisent pas correctement la réalité physique. Ainsi, dans le cas de représentations cubiques ou cylindriques, plusieurs points de l'espace peuvent représenter la même couleur (noir ou blanc) malgré des coordonnées différentes, ce qui n'est pas correct, contrairement aux représentations coniques et sphériques qui sont des espaces adaptés. Suivant la logique, on pourrait ajouter la représentation en demi-sphère pour le modèle HSV.

### 5.3.1.3 Classification dans l'espace de mesure adapté

Une fois défini un espace de calcul adapté au modèle de couleur considéré, il devient possible de classifier des données en mesurant les distances directement dans cet espace. Nous traitons ici le cas du modèle HSV, en utilisant un espace conique. Nous proposons d'abord une première solution traitant les coordonnées de façon séquentielle, avec une méthode adaptée pour la coordonnée polaire, puis une méthode traitant les trois coordonnées ensemble directement dans l'espace de représentation.

#### Classification séquentielle dans l'espace HSX

Afin de pouvoir utiliser la teinte pour la classification, il faut que celle-ci ait un sens, ce qui n'est pas le cas lorsque la saturation est nulle. Une solution peut alors être de traiter séparément les régions dites "achromatiques", où H n'a pas de sens, et les régions "chromatiques" (les "vraies" couleurs) où il en a un. Les régions achromatiques concernent les couleurs où la saturation est faible, c'est-à-dire le noir, le blanc et toutes les nuances de gris, mais aussi celles où la valeur est faible (couleur sombre proche du noir). C'est l'approche proposée dans [51], qui utilise le modèle HSI, ou encore dans [42] avec le modèle HSL. Dans ce cas, les régions achromatiques comprennent celles où l'intensité est faible (couleur proche du noir) ou forte (couleur proche du blanc), comme représenté sur la Figure 5.3.

Pour cela des seuils sont fixés arbitrairement pour distinguer les deux types de régions. Les régions achromatiques peuvent être traitées avec de nouveaux seuils si l'on souhaite distinguer le blanc, le noir et le gris. Quant aux régions chromatiques, elles peuvent être classifiées en utilisant uniquement la teinte avec un traitement adapté aux données périodiques, comme ceux proposés dans la Section 5.3.1.1. Cette solution permet de traiter une composante à la fois, ce qui était utile lorsque les ressources de calcul étaient trop limitées pour des approches par *clustering*. La difficulté majeure réside alors dans le choix des seuils pour les différentes régions.



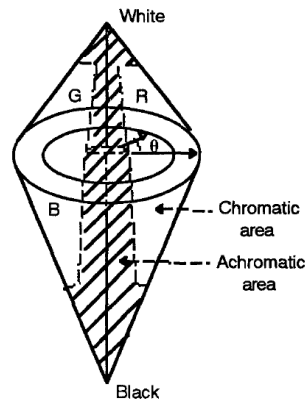


FIGURE 5.3 – Illustration du découpage de l'espace HSI en régions chromatiques et achromatiques (source : [51]).

### Classification $k$ -moyennes dans l'espace HSV conique

L'approche la plus directe pour classifier des données en HSV consiste à appliquer un *clustering* directement sur les coordonnées HSV des pixels de l'image, par exemple avec la méthode des  $k$ -moyennes. Or, comme détaillé dans la Section 5.3.1.2, ces coordonnées ne sont pas des coordonnées cartésiennes classiques, donc il est nécessaire de se placer dans un espace adapté, comme le cône pour le modèle HSV, ce qui revient à utiliser comme grandeurs les coordonnées définies dans l'équation 5.8. Il est alors possible de mesurer les distances directement dans cet espace, avec une distance classique comme la distance euclidienne (norme  $\mathcal{L}_2$ ), ou la distance de Manhattan (norme  $\mathcal{L}_1$ ) qui est plus adaptée [2].

$$\text{Avec } (S, V) \in [0, 1]^2 \text{ et } H \in [0, 2\pi], \begin{cases} x = SV \cos H \\ y = SV \sin H \\ z = V \end{cases} \quad (5.8)$$

Avec ce modèle, il n'est pas possible de donner plus ou moins de poids à H exclusivement, puisque les coordonnées impliquant H impliquent aussi S et V. Cela dit, il est possible de donner un poids moins important à V, et donc plus important aux autres, ou inversement, en jouant sur le poids de la coordonnée  $z$  relativement aux deux autres, ce qui revient à aplatir ou étirer le cône. En l'aplatissant, les distances suivant la coordonnée  $z$  ont alors moins d'impact que celles suivant  $x$  et  $y$ , et inversement. La composante H étant liée à la chromaticité, contrairement à V qui est dédiée à la luminosité, cette solution permet de pondérer ces deux aspects.

Cette approche simple est pourtant absente de la littérature, à notre connaissance en tout cas. Il est possible qu'elle n'ait pas été abordée du fait des limites des ressources de calcul au moment du développement des espaces de couleurs comme le HSV, puis oubliée au profit d'autres solutions arrivées ensuite. Elle est similaire à celle proposée initialement en se basant sur les coordonnées trigonométriques (cf. Section 5.3.1.1), à la différence qu'ici les 3 grandeurs sont combinées en tenant compte des relations issues de leur définition, en plus de l'aspect périodique du H. En fait, l'approche initiale combinant la teinte et les autres grandeurs de manière indépendante revient à considérer que la teinte évolue sur un cercle unité tandis que les deux autres grandeurs évoluent chacune dans une autre dimension indépendante, ce qui nécessite un espace à 4 dimensions.

### 5.3.2 Expérimentations et résultats dans l'espace HSV

Afin de valider notre approche pour l'utiliser dans une chaîne de traitement plus large, nous avons mené des tests qualitatifs sur un jeu d'images réelles, en utilisant l'espace HSV. Dans cette section nous présentons ce jeu de données, le périmètre des tests et les résultats obtenus.

### 5.3.2.1 Protocole expérimental

#### Données de test

Nous avons évalué nos approches sur trois séquences du jeu d'images aériennes *Aeroscapes* [28]<sup>1</sup>. Ces données sont présentées dans la Section 1.2.2.3, et des extraits de ces séquences sont donnés dans la Figure 1.11. Pour nos tests, nous avons extrait les personnes grâce aux masques de segmentation associés, afin de les décomposer en parties pour une tâche de ré-identification (cf. Chapitre 6). Pour nos tests qualitatifs, nous avons en particulier utilisé une image de la séquence 040000, qui a l'intérêt d'être suffisamment résolue et difficile pour la tâche. En effet, elle contient plusieurs couleurs assez proches (plusieurs variations de bleu plus ou moins lumineux), ainsi que du gris, du noir et du blanc, ce qui est intéressant dans l'espace HSV. La Figure 5.4 donne plusieurs visualisations de cette image par composantes. Dans nos tests, nous utilisons soit les pixels issus du masque de segmentation de la personne, soit l'ensemble des pixels du rectangle englobant celle-ci.



FIGURE 5.4 – Visualisation des composantes HSV d'un extrait d'une image du jeu de données *Aeroscapes* (à gauche : pixels du masque uniquement, à droite : rectangle englobant). Pour la composante H, le S et le V sont ignorés, ou réglés à 1, afin de visualiser la teinte "pure" (`cmap 'hsv'` dans `python`). Pour la composante S, le H et le V sont ignorés, et les valeurs vont du blanc pour S=0 au noir pour S=1 (`cmap 'binary'`). Pour la composante V, le H et le S sont ignorés, et les valeurs vont du noir pour S=0 au blanc pour S=1 (`cmap 'gray'`). À noter que la valeur de H n'est pas toujours définie suivant les valeurs de S et V, et que la valeur de S n'est pas toujours définie suivant la valeur de V (cf. Figure 5.2).

#### Méthodes évaluées

L'approche générale utilisée pour toutes nos évaluations est basée sur un *clustering* par la méthode des *k*-moyennes [26], adaptée suivant les approches présentées dans la Section 5.3.1. Différentes méthodes ont été évaluées sur l'image de test, à des fins de comparaison uniquement pour certaines (marquées par un (c)) :

- (c) les composantes RGB,
- (c) l'image en niveaux de gris,
- la composante H seule, selon l'approche décrite dans la Section 5.3.1.1, avec plusieurs variantes :
  - (c) sans tenir compte de la périodicité,
  - en tenant compte de la périodicité grâce à un encodage trigonométrique,
  - en tenant compte de la périodicité en la considérant comme un angle,
- les composantes HSV, selon les approches de la Section 5.3.1.3, avec plusieurs variantes :
  - (c) vues comme des coordonnées cartésiennes classiques, sans tenir compte de la périodicité de H ni de l'interdépendance entre S et V,
  - en considérant H comme un angle et S et V comme des coordonnées cartésiennes classiques,
  - avec un traitement séquentiel, et en tenant compte de la périodicité pour le H en le considérant comme un angle,
  - en considérant H comme un angle et en tenant compte de l'interdépendance entre S et V, en se plaçant dans l'espace conique.

1. <https://github.com/ishann/AeroScapes>

Des variations de paramètres sont alors possibles selon les méthodes. L'impact des paramètres suivants a été évalué :

- le nombre de *clusters* (ou de centroïdes) pour les  $k$ -moyennes : de 3 à 6 ;
- l'initialisation des centroïdes pour les  $k$ -moyennes : aléatoire ou fixe (pour le cône : 3 sur l'axe vertical et  $k-3$  sur le cercle de base, cf. Figure 5.7a) ;
- le critère de distance maximale pour l'arrêt des itérations de l'algorithme des  $k$ -moyennes ;
- les poids donnés aux différentes composantes dans le HSV : normalisation (même poids à tous) ou pas, poids de 0.5, 1 ou 2 pour la composante H (et 1 pour les autres) ;
- les seuils sur S et V pour les régions achromatiques dans le traitement séquentiel.

Concernant la normalisation, il est à noter qu'en général, les valeurs des composantes sont encodées sur un nombre de valeurs numériques identique, en fonction de la taille mémoire allouée (256 valeurs par exemple pour un encodage sur 8 bits). Or, la composante H étant angulaire, on a pris l'habitude de la coder sur des multiples ou diviseurs de 360, comme 180 sur 8 bits, sans utiliser les valeurs supérieures (de 180 à 255), tandis que les autres composantes du HSV couvrent tout l'intervalle. Il est donc utile de procéder à une étape de normalisation afin de donner un poids identique à chacune des composantes, plutôt que de les utiliser telles quelles. Un poids différent peut alors être donné à une composante pour modifier son impact, comme décrit précédemment.

### 5.3.2.2 Résultats et analyse

Nous avons uniquement mené des tests qualitatifs et présentons les résultats obtenus sur l'image de référence ainsi que sur d'autres images de la même séquence et d'une autre séquence du jeu de données *Aeroscapes*. Nous donnons ici une comparaison des résultats obtenus avec les différentes méthodes, puis nous étudions l'impact de la pondération des composantes et de l'initialisation des centroïdes dans le HSV, avant de donner des résultats sur les séquences avec la meilleure méthode retenue.

#### Comparaison des différentes méthodes

Tout d'abord, nous avons pu constater qu'utiliser uniquement la composante H conduisait à des segmentations peu pertinentes, ne permettant pas de distinguer clairement les parties de couleurs différentes dans l'image test si ce n'est l'ensemble casquette+visage, les baskets, et partiellement l'étiquette de la casquette et le dossard sur la veste. Les autres parties importantes (la veste bleu clair, le pantalon bleu, les cheveux, le fond) se retrouvent mélangées dans une seule grande classe, avec des nuances sur certaines régions lorsque l'on ajoute des classes (dans les cheveux et le fond, sur le bras droit). En fait et par définition, cette composante seule ne permet pas de distinguer des couleurs claires proches du blanc (pour la veste) ou foncées proches du noir (pour les cheveux), les réduisant à leur teinte (bleue) même si elle n'est pas significative. La visualisation de la composante de teinte dans la Figure 5.4 montre bien cela sur l'image d'entrée, où l'on peut constater qu'il est en fait impossible de distinguer ces parties. De la même façon, le HSV séquentiel ne permet pas de séparer ces couleurs, puisqu'il utilise également le H seul pour les régions chromatiques, et le seuillage des couleurs achromatiques ne permet pas de séparer ces parties. De plus, il est très difficile de fixer les seuils pour cette méthode.

Il est donc nécessaire d'utiliser les autres composantes pour obtenir une segmentation permettant de distinguer toutes les couleurs. Les sorties obtenues avec deux des méthodes proposées pour cela sont reproduites et comparées avec deux méthodes plus simples dans la Figure 5.5. Nous représentons les sorties en donnant à chaque pixel la couleur du centroïde de sa classe, ce qui revient à appliquer un échantillonnage des couleurs. Nous donnons à la fois les résultats obtenus à partir des pixels correspondant au masque de segmentation uniquement et à partir de l'ensemble des pixels du rectangle englobant. Ne pas avoir à traiter le fond permet de donner plus d'importance aux autres classes, ce qui peut être utile pour des valeurs de  $k$  faibles, mais il suffit d'autoriser plus de classes pour pouvoir segmenter correctement le fond et les autres classes. Ainsi, les deux options donnent des résultats équi-

valents pour les valeurs de  $k$  considérées, ce qui permet de prendre en entrée uniquement des détections sous forme de rectangles englobants, sans nécessiter de segmentation en amont.

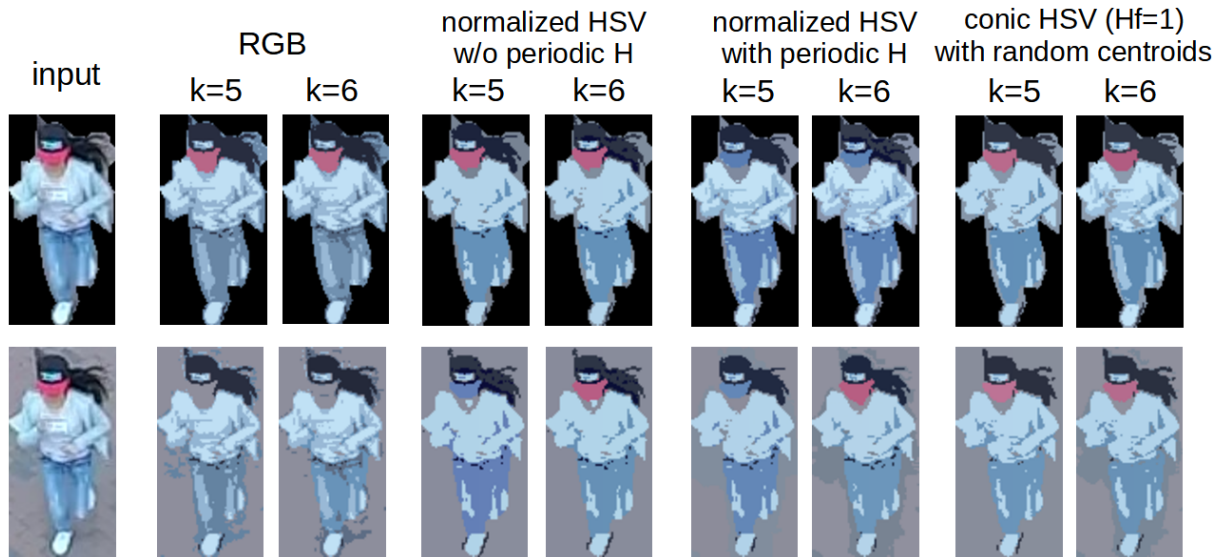


FIGURE 5.5 – Sorties de segmentation avec les différentes méthodes proposées (en attribuant à chaque pixel la couleur du centroïde de sa classe).

En comparant les différentes méthodes, on constate d'abord que les résultats obtenus avec le RGB sont satisfaisants sur le masque, tandis qu'ils ne distinguent pas la couleur de la casquette pourtant très différente sur le rectangle englobant, même avec  $k=6$ , ce qui renforce l'intérêt pour des modèles prenant mieux en compte la teinte. Les résultats obtenus avec le HSV non périodique et le HSV périodique sont également satisfaisants, le principal défaut étant à nouveau la casquette rouge qui se retrouve dans la même classe que le pantalon bleu, pourtant éloignés en termes de teinte, l'algorithme se focalisant plus facilement sur des distinctions dans la couleur des cheveux. Il serait alors utile de donner un poids plus important à la composante de teinte, qui permet bien de séparer la casquette d'après les résultats préliminaires sur H seul, ou d'initialiser différemment les centroïdes (ce qui est fait de manière aléatoire ici). En revanche, le HSV conique permet de bien séparer ces classes dès  $k=5$ , que ce soit avec une initialisation des centroïdes aléatoire ou donnée (cf. la Figure 5.6).

### Impact de la pondération des composantes dans le HSV

Afin de faire varier le poids à la teinte (ou plutôt de la chromaticité, par rapport à la luminosité), nous avons multiplié les coordonnées  $x$  et  $y$  dans le cône par un facteur  $H_f$ , ce qui revient à aplatir ou étirer le cône. Nous avons étudié l'impact de ce facteur en étudiant les valeurs 0.5, 1 et 2. Les sorties de la segmentation sont données dans la Figure 5.6. En analysant ces sorties, on constate qu'il y a plus de petits éléments qui ressortent lorsque l'on diminue le poids de la teinte, notamment sur le pantalon et la veste, qui sont alors dus aux variations de luminosité ou de saturation. Sur la version à partir du masque de segmentation, où il y a nettement moins de pixels pour le fond, on ne fait plus la différence entre le fond gris et le pantalon bleu lorsque  $H_f=0.5$ . Ceci peut s'expliquer par le fait qu'en réduisant  $H_f$ , les couleurs sont rapprochées du gris et la comparaison se fait moins selon la teinte et la saturation mais davantage selon la valeur.

La valeur de  $H_f$  a donc bien un impact sur la classification et sur l'importance donnée à la chromaticité ou la luminosité. Modifier cette valeur permet alors de donner plus d'importance à l'une ou l'autre. Suite à nos expérimentations, nous avons choisi de fixer  $H_f=1.5$  pour la segmentation des séquences, qui est utilisée dans la chaîne de ré-identification du Chapitre 6.

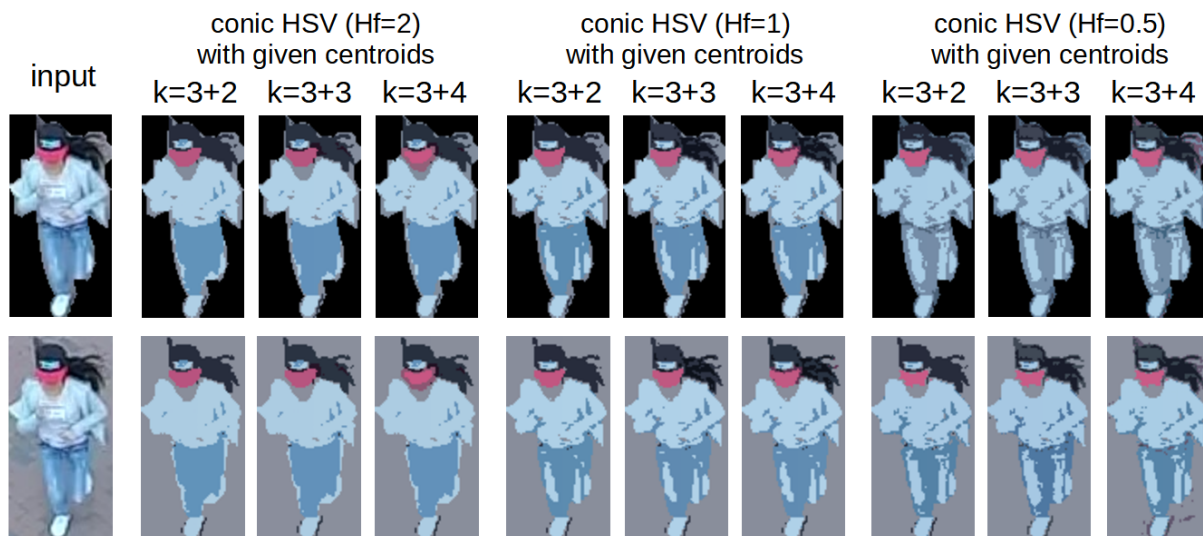


FIGURE 5.6 – Impact du poids de la composante H dans la segmentation dans l'espace HSV conique : résultats sur l'image test en faisant varier Hf et k, avec une initialisation des centroïdes donnée.

### Impact de l'initialisation des centroïdes pour le HSV

Afin d'orienter la classification par  $k$ -moyennes sur des couleurs plus variées, nous avons également évalué l'impact de l'initialisation des centroïdes, pour la composante de teinte en particulier. L'idée est de forcer l'algorithme à utiliser des centres éloignés, ce qui n'est pas forcément le cas en initialisant de façon aléatoire parmi les données existantes, où l'initialisation est influencée par la distribution des données. Cette variation a été évaluée avec le HSV conique, en choisissant des centroïdes initiaux sur l'axe vertical et sur le cercle de base du cylindre. Sur l'axe, on considère 3 centroïdes de saturation nulle, avec les valeurs  $V=0, 0.5$  ou  $1$ , tandis que sur le cercle on considère  $k-3$  centroïdes équirépartis, comme le montrent les Figures 5.7a et 5.8b.

Les résultats sur l'image de test sont très similaires aux précédents (cf. Figures 5.6 et 5.5) : pour  $k=5$  ils sont quasiment identiques, et pour  $k=6$  on trouve deux classes différentes pour les couleurs des cheveux avec les centroïdes fixés, tandis que la nouvelle classe est soit sur la veste soit sur le fond avec les centroïdes aléatoires. Aucune des deux options ne permet donc de séparer d'autres couleurs comme celle du visage ou des chaussures ici (ces objets sont cependant séparables en traitant les composantes connexes). Il serait intéressant d'effectuer des évaluations sur d'autres données pour vérifier la capacité d'apporter de la diversité. Par ailleurs, une autre option possible pour donner moins de poids aux données très représentées et plus de poids à celles peu représentées serait de modifier la distribution des données en entrée, par exemple en ne tenant pas compte des valeurs déjà beaucoup représentées ou proches d'une valeur déjà beaucoup représentée.

## 5.4 Conclusion

Dans ce chapitre, nous avons proposé une méthode de segmentation d'images basée sur la couleur. Pour cela, nous avons d'abord mis en évidence l'importance du choix de l'espace de couleurs et de l'utilisation de méthodes adaptées pour mesurer des écarts de couleurs dans ces espaces, en particulier dans les espaces polaires de type HSX qui sont souvent utilisés. La méthode que nous proposons consiste à se placer dans un espace géométrique adapté, c'est-à-dire traduisant bien la définition des composantes, de façon à ce que les distances mesurées traduisent bien les écarts perceptuels définis par le modèle. Pour le modèle HSV, nous utilisons l'espace conique en traduisant en coordonnées cartésiennes les coordonnées polaires, se rapprochant de la définition des modèles XYZ,  $L^*a^*b^*$  ou  $L^*u^*v^*$ , dont la traduction en



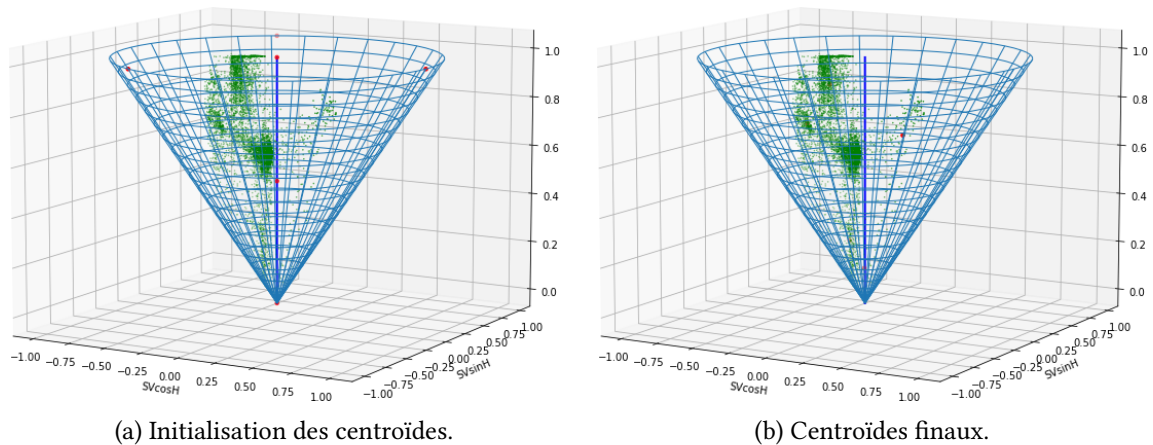


FIGURE 5.7 – Visualisation 3D de la segmentation dans le cône HSV : visualisation des valeurs des pixels (en vert) et des centroïdes (en rouge), sur l’image de test.

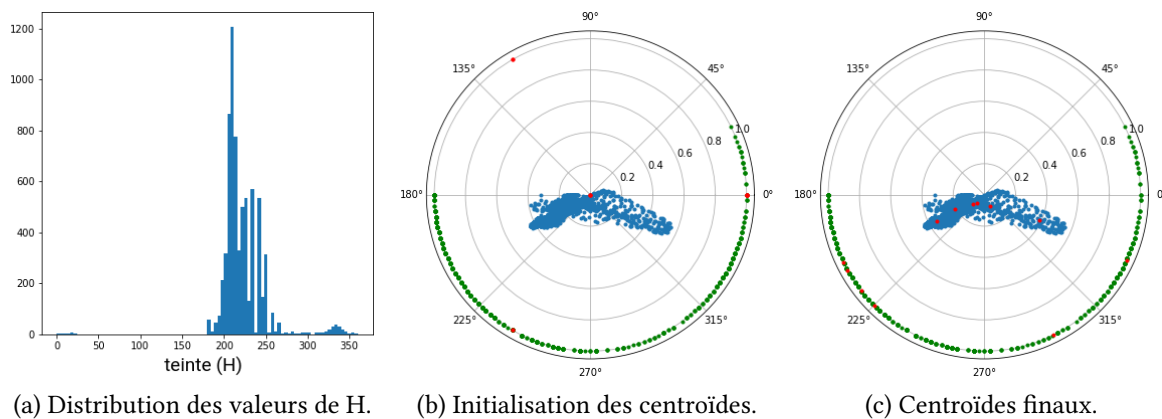


FIGURE 5.8 – Visualisation 2D de la segmentation sur les valeurs de teinte : répartition des valeurs de H des pixels (en vert sur les cercles unité, en bleu par vue de dessus du cône HSV) et des centroïdes (en rouge), sur l’image de test.

polaire est plutôt un modèle HCL en considérant la chrominance absolue.

Notre solution se base sur un *clustering* selon la méthode des  $k$ -moyennes, ce qui permet de s’adapter aux données contrairement à un simple seuillage, tout en gardant une solution simple. Nous l’avons évaluée sur plusieurs séquences d’images aériennes de suivi de personnes, afin d’obtenir des décompositions selon la couleur des vêtements. Les résultats sont très satisfaisants, ce qui permet de vérifier que la couleur est une caractéristique très utile pour la segmentation d’images, voire suffisante selon le cas d’usage. Ils permettent alors d’envisager une exploitation des parties obtenues pour des tâches de compréhension de scènes, après un traitement pour extraire les composantes connexes et éventuellement un post-traitement par opérations morphologiques. En particulier, cette décomposition peut être utile pour décrire une personne et la reconnaître ultérieurement, à partir d’un ou quelques exemples seulement, approche que nous avons utilisée dans la chaîne de traitement de ré-identification proposée dans le Chapitre 6 de ce manuscrit.

Les espaces HSX ont l’avantage de permettre de séparer les composantes de chromaticité et de luminosité, ce qui peut ensuite être exploité pour ne pas tenir compte de la luminosité typiquement, afin d’avoir des approches robustes à des variations de ce paramètre. Bien que beaucoup de travaux utilisent ces modèles de couleurs pour la segmentation, aucun à notre connaissance n’a cherché à évaluer cette capacité, exploitant tous les trois composants sans traitement spécifique. Dans notre solution sur le

HSV, il est possible de donner plus d'importance à la chromaticité ou à la luminosité en jouant sur la pondération des composantes, ce qui revient à aplatir ou allonger le cône HSV. Nous avons pu vérifier ce comportement dans nos évaluations, et également observer une invariance à la luminosité sur la décomposition, sur les séquences utilisées. Afin de ne pas tenir compte de la luminosité pour la ré-identification, l'idée est alors d'utiliser une pondération "normale" pour la décomposition de l'objet en parties, et un poids plus important pour la chromaticité pour la comparaison des parties entre images (*i.e.*, moins important pour la luminosité).

Cependant, la capacité des modèles HSX à séparer la chromaticité et la luminosité n'a pas été évaluée précisément. Il serait donc intéressant d'aller plus loin dans l'étude, avec une évaluation sur des images présentant des variations de l'une ou l'autre composante par exemple. Il serait bon aussi d'élargir l'évaluation à davantage de données, et de mener une étude quantitative sur un jeu de données complet, afin de comparer plus précisément les approches et paramétrages. Par ailleurs, d'autres modèles de couleurs plus évolués pourraient être utilisés, comme les espaces  $L^*a^*b^*$  ou  $L^*u^*v^*$ , et d'autres méthodes de segmentation également, comme l'algorithme *mean shift* ou le partitionnement de graphes. De plus, pour une segmentation plus robuste, il peut être pertinent de combiner la couleur à d'autres caractéristiques comme la texture ou encore des a priori géométriques sur la forme ou les relations spatiales entre objets. Enfin, il faut noter que les espaces de couleurs utilisés couramment sont destinés à modéliser la perception humaine, tandis que les capteurs image peuvent potentiellement acquérir un plus large spectre colorimétrique, et que d'autres modélisations et traitements des couleurs sont possibles. Suivant l'application, il peut être utile de modéliser fidèlement la perception humaine ou d'exploiter un spectre plus large, en traitant le signal brut en sortie du capteur ou avec d'autres modélisations.



## **Partie III**

# **Application à la ré-identification**



## Chapitre 6

# Ré-identification par appariement de parties

---

6.1	Introduction . . . . .	113
6.1.1	Contexte et positionnement des travaux . . . . .	113
6.1.2	Applications et jeux de données . . . . .	115
6.1.3	Axes de recherche de la littérature . . . . .	117
6.1.3.1	De la reconnaissance fine à la ré-identification . . . . .	117
6.1.3.2	Ré-identification de personnes . . . . .	118
6.2	Décomposition en parties et comparaison par appariement des parties . . . . .	119
6.2.1	Approche proposée . . . . .	119
6.2.2	Expérimentations et résultats . . . . .	122
6.2.2.1	Données et protocole expérimental . . . . .	122
6.2.2.2	Résultats et discussion . . . . .	123
6.3	Conclusion . . . . .	124

---

Dans ce chapitre, nous explorons le problème de la ré-identification de scènes ou objets, afin d'améliorer une solution développée précédemment par l'entreprise Magellium, notamment pour le cas de la ré-identification de personnes dans des images aériennes. Suite à une étude approfondie de la solution existante, nous avons mené une refonte de celle-ci afin de mieux traiter chaque étape et en particulier la modélisation de la configuration spatiale. Pour cela, nous proposons une approche basée sur l'appariement des parties des objets, selon trois étapes : décomposition de la scène, description des parties par différentes caractéristiques, appariement et comparaison des parties. Cette solution peut ensuite être utilisée pour un apprentissage de prototypes à partir de quelques données seulement, tout en étant robuste à des variations de certaines caractéristiques. Dans nos travaux, nous avons d'abord exploité trois caractéristiques basiques : la couleur, la taille et la position des parties. Puis nous avons exploré l'utilisation d'un descripteur de configuration spatiale, ce qui inclue une description de la forme. Nous avons évalué notre solution sur des séquences d'images aériennes, ce qui a permis de montrer l'intérêt de l'approche et l'importance du choix des caractéristiques pour l'appariement et la comparaison, en fonction du cas d'usage.

## 6.1 Introduction

### 6.1.1 Contexte et positionnement des travaux

Afin d'améliorer une solution développée précédemment par l'entreprise Magellium, nous avons exploré le problème de la ré-identification de scènes ou d'objets, notamment la ré-identification de

personnes. Cette tâche récente en vision par ordinateur a pour but de reconnaître un élément d'intérêt qui a été vu une ou quelques fois auparavant, dans une vidéo par exemple. Elle s'apparente donc à la reconnaissance à partir de peu d'exemples, qui est une tâche ancienne, à la différence que la ré-identification ne cherche pas à reconnaître une classe générique d'objet, mais à vérifier s'il s'agit du même exemple d'une classe générique (la même personne par exemple), dont on dispose de quelques vues, ce qui donne des classes individuelles plus précises. Elle se distingue par contre de l'identification à partir de données biométriques, où l'on cherche à vérifier qu'il s'agit du même exemple mais avec une acquisition contrôlée, sur des données spécifiques (empreinte, iris, etc.).

Lorsque l'on parle de ré-identification, il faut en fait distinguer la ré-identification court terme et la ré-identification long terme. La ré-identification long terme est parfois appelée "recherche", rejoignant la tâche de recherche dans une base (CBIR), où l'on cherche les éléments les plus proches dans la base (*ranking*), qui est en réalité très proche de la tâche de reconnaissance (classification) [22]. Par opposition, la ré-identification court terme rejoint quant à elle la tâche de "suivi", où l'on cherche à identifier la cible dans des images successives d'une séquence, donc avec un mouvement limité entre chaque image. La Figure 6.1 illustre la différence entre ces deux tâches sur des personnes, où certains attributs caractéristiques à court terme (la tenue vestimentaire, les accessoires) ne le sont plus à long terme. La ré-identification à long terme est donc plus complexe puisqu'elle a moins d'attributs sur lesquels s'appuyer. En résumé, on peut donner la catégorisation des tâches suivante :

- suivi : on souhaite suivre une cible dans une séquence d'images où elle est toujours présente ;
- ré-identification court terme : on souhaite savoir si une cible est la même qu'une autre vue peu de temps avant (dans la même séquence typiquement) ;
- ré-identification long terme ou recherche dans une base : on souhaite savoir si une cible a déjà été vue, *i.e.*, si elle est dans la base ;
- recherche d'images similaires dans une base (CBIR) : on cherche les éléments les plus proches dans la base.

Bien que toutes ces tâches entrent dans nos cas d'usage, nos expérimentations se sont focalisées sur la ré-identification à court terme.



FIGURE 6.1 – Illustration des tâches de ré-identification à court-terme et long-terme (d'après [2]).

Étant donné qu'il s'agit de construire un modèle de cible afin de la reconnaître, la ré-identification s'apparente à une tâche d'apprentissage, qui est alors faiblement supervisé (*few-shot/one-shot learning*) puisqu'on dispose de peu d'exemples. Elle est également en lien avec l'apprentissage continu ou incrémental, dans le cas où l'on a la possibilité d'ajouter de nouveaux exemples après un premier apprentissage. Mais la problématique principale pour la ré-identification est sans doute celle de l'appariement, incontournable dès lors que l'on traite des images dont les conditions d'acquisition ne sont pas contrôlées, en particulier le point de vue ou la luminosité. Cette problématique nécessite d'utiliser

des approches robustes, soit en détectant et modélisant les variations afin d'en tenir compte (la pose des objets en 3D par exemple), soit en utilisant des caractéristiques invariantes, liées à la nature des objets, comme la couleur des vêtements pour la ré-identification de personnes. Dans nos travaux, nous n'avons pas abordé le problème de la modélisation du point de vue, mais nous proposons une approche basée sur une combinaison de caractéristiques peu variables.

Enfin, dans le cas de systèmes opérationnels, comme ceux sur lesquels travaille l'entreprise Magellium, la problématique de temps réel est également importante, la chaîne développée ayant d'ailleurs été appelée "apprentissage et classification à la volée". En effet, le modèle est construit directement ("*online*" ou "à la volée") lors du traitement d'une vidéo, image par image d'apprentissage, et la reconnaissance s'effectue directement aussi, à partir de sélections par un utilisateur ou de détections automatiques. Cet aspect n'a pas été abordé explicitement dans nos travaux, mais la solution proposée permet d'envisager un traitement en temps réel.

Face à tous ces défis, la modélisation de la configuration spatiale peut avoir un intérêt, que ce soit pour reconnaître des scènes ou des objets qui ont un agencement particulier, ou pour apparier des scènes ou des objets avec des configurations similaires afin de les comparer. Ainsi, la solution Magellium s'appuie sur une décomposition des objets en parties et une description de ces parties basée sur plusieurs caractéristiques, dont leur configuration spatiale. Cependant, cette solution n'a pas entièrement donné satisfaction, ce qui a conduit à vouloir l'améliorer, en particulier la description des relations spatiales. Après une étude approfondie de la solution actuelle, nous avons proposé une refonte de celle-ci afin de mieux traiter chaque étape et en particulier la modélisation de la configuration spatiale, dans le but d'obtenir une preuve de concept pour cette approche.

### 6.1.2 Applications et jeux de données

La ré-identification consiste à reconnaître un objet précis ou une scène précise à partir de quelques exemples, pour n'importe quel type d'objet ou de scène. Néanmoins, seuls deux cas d'usage ont un véritable intérêt et sont couramment traités dans la littérature : la ré-identification de personnes et la ré-identification de véhicules, avec de nombreux jeux de données disponibles. Les applications couvrent principalement le domaine de la sécurité : le contrôle d'identité pour les personnes, la reconnaissance et le suivi d'individu ou de véhicule suspect, avec un réseau de caméras typiquement, ou encore la recherche dans une base existante. Ces tâches peuvent concerner deux cadres différents : la vidéo-surveillance avec des caméras au sol, avec une vue de face ou légèrement sur-élevée, ou la surveillance par des capteurs aériens (drones, avions, ballons...), avec une vue oblique ou de dessus (vue zénithale). La ré-identification de scènes peut quant à elle être utilisée pour la localisation basée vision.

Un paramètre important pour la ré-identification est la résolution spatiale des images, car elle définit le type d'objets, de parties d'objets ou de détails qui peuvent être exploités pour distinguer les scènes ou les objets. À basse résolution, seules les éléments de grande taille sont discernables, ce qui rend la tâche compliquée voire impossible selon le cas ; à une résolution intermédiaire, davantage d'éléments et de caractéristiques peuvent être exploités ; tandis qu'à haute résolution, il est possible de s'appuyer sur des détails pour obtenir des indices caractéristiques. Par exemple, pour la reconnaissance de véhicule, à basse résolution on ne pourra exploiter que la forme et la couleur des parties, tandis qu'à haute résolution des détails comme un logo, une inscription ou la position d'un élément deviendront exploitables. Pour la reconnaissance de personnes, à basse résolution on ne pourra exploiter que les parties principales comme la tête, le torse et les jambes ; à moyenne résolution on pourra décomposer les parties du corps avec une approche de segmentation ("*human parsing*", cf. Figure 6.2) ; et à haute résolution on pourra reconnaître les motifs ou écritures sur un vêtement, un tatouage, un petit accessoire comme un bracelet, des lunettes, etc.

Le cas d'usage que nous visons est celui de la ré-identification de personnes à court terme, à partir d'imagerie aérienne en vue oblique, à basse ou moyenne résolution, ce qui permet d'avoir une décomposition en parties plus ou moins fine. Un recensement des jeux de données pour la ré-identification



FIGURE 6.2 – Exemple de segmentation des parties du corps ("human parsing") sur une image test (d'après [16]).

de personnes est disponible dans [2]. On constate que de nombreux jeux de données existent pour la vidéo-surveillance, mais assez peu pour l'imagerie aérienne. La Figure 6.4 reprend le tableau de synthèse de [2] en filtrant sur les données d'imagerie aérienne destinées à la ré-identification. Au final, seuls quatre jeux de données correspondent à ces critères : MRP [15], DRone HIT [9], PRAI-1581 [33] et P-DESTRE [2]. Ces jeux de données ont des résolutions variables, avec des altitudes entre 5m et 60m et des capteurs légers, comme détaillé dans la Figure 6.3 pour le jeu de données P-DESTRE. D'autres jeux de données aériennes de suivi de personnes sont également utilisables, comme le jeu de données *Aeroscapes* [20] qui contient des segmentations des éléments de la scène.

Cependant, ces jeux de données ne contiennent pas de segmentation des parties, ce qui nécessite d'utiliser une étape supplémentaire pour pouvoir les extraire. Pour cela, plusieurs approches basées sur la détection de pose ont été proposées, comme celle de [16] dont des résultats sont présentés dans la Figure 6.2, mais leur utilisation pour des images aériennes moins résolues reste à vérifier. La solution utilisée dans nos expérimentations est une décomposition basée sur la couleur, selon l'approche présentée dans le Chapitre 5. Nous avons ainsi pu obtenir plusieurs séquences segmentées par parties à partir du jeu de données *Aeroscapes*, comme détaillé dans la Section 1.2.2.3.



FIGURE 6.3 – Protocole d'acquisition du jeu de données P-DESTRE (d'après [2]).

Dataset	Format	Task					Identities	Bound. Box	Height (m)
		Detection	Tracking	ReID	Search	Action Rec.			
MRP	Video	✓	✓	✓	✗	✗	28	4K	< 10
PRAI-1581	Still	✗	✗	✓	✗	✗	1,581	39K	[20, 60]
DRone HIT	Still	✗	✗	✓	✗	✗	101	40K	25
P-DESTRE	Video	✓	✓	✓	✓	✓	269	> 14.8M	[5.5, 6.7]

FIGURE 6.4 – Synthèse des jeux d'images aériennes pour la ré-identification (d'après [2]).

### 6.1.3 Axes de recherche de la littérature

La ré-identification consiste à reconnaître n'importe quelle scène ou n'importe quel objet précis vu une ou quelques fois seulement (dans plusieurs trames d'une séquence vidéo typiquement) et possible-ment dans des conditions d'acquisition différentes (pose des objets, point de vue, capteur, luminosité...). Il s'agit donc d'un problème de reconnaissance à grain fin à partir de peu de données d'entraînement, comme illustré par la Figure 6.5, qui est un problème étudié depuis les débuts de la reconnaissance d'objets, alors que l'on avait peu de données pour entraîner les modèles. Ce sujet est lié à de nombreuses autres problématiques : l'apprentissage à partir de peu de données (*few-shot learning*), la reconnaissance multi-vues, l'estimation de pose, ou encore l'apprentissage incrémental lorsque l'on autorise l'ajout ultérieur de nouvelles données d'apprentissage.

#### 6.1.3.1 De la reconnaissance fine à la ré-identification

Considérant qu'il s'agit d'une tâche de reconnaissance à partir de peu d'exemples, les approches originelles basées sur des descripteurs de caractéristiques des régions ou sur des points d'intérêts constituent une base intéressante pour ce problème, en utilisant des caractéristiques suffisamment fiables et robustes pour le cas d'usage. Différents types de descripteurs peuvent être utilisés pour cela : des descripteurs de caractéristiques concrètes comme la couleur, la texture, la forme, le contour, la structure (le squelette) ou la position (du barycentre ou d'autres points d'intérêt), mais aussi des descripteurs plus abstraits basés notamment sur des variations d'intensité autour de points d'intérêt (MSER, HOG, SIFT, SURF...). La recherche s'est notamment attachée à développer des descripteurs invariants aux différentes transformations pouvant affecter les objets [31], ainsi que des méthodes de détection de régions stables sur lesquelles calculer ces descripteurs [27, 13]. [10] est alors un des premiers à combiner ces deux aspects pour la reconnaissance d'objets, utilisant des descripteurs de forme, de couleur et de texture sur une décomposition hiérarchique des objets en régions.

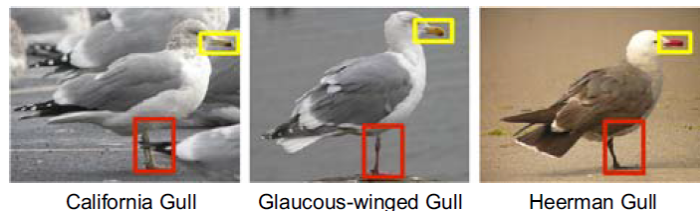


FIGURE 6.5 – Illustration de la tâche de reconnaissance à grain fin (source : [30]).

D'autres approches ont ensuite été développées plus spécifiquement pour la reconnaissance à grain fin, plusieurs jeux de données de tailles plus importantes ayant été développés pour cette tâche (cf. Figure 6.5). Par exemple, [30] propose de rechercher les caractéristiques discriminantes à l'aide d'arbres de décision sur les descripteurs de régions, avec une sélection aléatoire afin d'éviter le sur-apprentissage. Plutôt que de traiter toutes les parties en même temps, [11] propose une approche séquentielle, de façon à utiliser la partie la plus discriminante à chaque étape, en s'inspirant de la vision humaine par saccades (cf. Figure 6.6). Ces méthodes utilisent respectivement des descripteurs SIFT et HOG pour leurs évaluations, mais elles s'emploient sur tout type de descripteur.

Avec l'essor des CNN et le développement de jeux de données toujours plus grands, de nombreuses approches ont été proposées pour cette tâche de reconnaissance fine [32, 17, 29, 7, 34]. Dans ces approches, l'invariance aux transformations est obtenue par construction pour la translation, par augmentation de données pour la rotation, les déformations et les variations de luminosité, et grâce à des architectures multi-échelles pour la variation d'échelle. Plusieurs travaux ont alors cherché à apporter de l'explicabilité dans ces approches, comme ProtoPNet [4] qui combine pour cela des preuves issues de plusieurs parties des objets en les associant à différents prototypes, dans la lignée des approches de



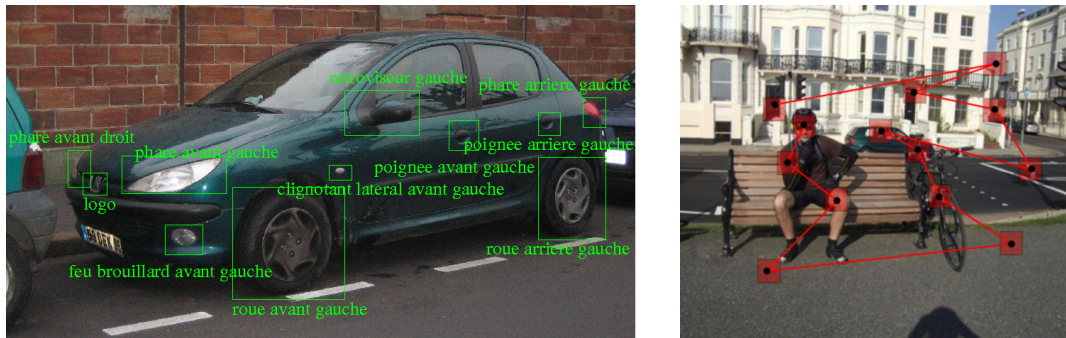


FIGURE 6.6 – Exemple de différents détails d’un objet et illustration du processus de vision par saccade (d’après [11]).

raisonnement à partir de cas (*case-based reasoning*). Par ailleurs, une autre approche similaire basée sur des réseaux de neurones est celle des réseaux à capsules (*capsule networks*) [12, 23], où l’idée est de reconnaître un objet en accumulant des preuves de la présence de ses parties avec les mêmes paramètres de pose. Cependant, tous ces modèles restent réservés aux cas où l’on dispose de quantités importantes de données d’apprentissage.

Enfin, d’autres approches ont été développées plus spécifiquement pour la ré-identification. Elles peuvent être classées en quatre catégories :

- les solutions orientées sur les caractéristiques, qui visent à élaborer des descripteurs de caractéristiques discriminants et invariants aux conditions d’acquisition, avec notamment des approches cherchant à apprendre automatiquement quelles caractéristiques utiliser suivant le cas ;
- les solutions par apprentissage de métrique, dans le but de minimiser la distance intra-classe et de maximiser la distance inter-classes ;
- les solutions basées sur le raisonnement à partir de cas (*case-based reasoning*), où l’on cherche à définir des prototypes pour pouvoir s’y comparer ;
- les solutions basées sur l’apprentissage profond.

### 6.1.3.2 Ré-identification de personnes

Pour la ré-identification de personnes, on distingue la ré-identification à court terme, dans des images proches d’une séquence typiquement, et la ré-identification à long terme, qui s’apparente plus à de la recherche dans une base, comme détaillé dans la Section 6.1.1.

Plusieurs états de l’art ou études comparatives sont disponibles dans la littérature. [25] s’intéresse à la détection, au suivi et à la ré-identification de personnes par un réseau de caméras, et propose ainsi une première étude approfondie et un état de l’art des techniques dédiées à cette tâche. [8] s’intéresse particulièrement la ré-identification et dresse une évaluation comparative complète des méthodes selon un protocole commun, en utilisant diverses caractéristiques, métriques et jeux de données. [1] se focalise sur les approches basées sur l’apprentissage profond, et [18] sur le problème de ré-identification à long terme. [2] met en évidence les six principaux défis de la ré-identification de personnes, qui sont (cf. Figure 6.7) : la faible résolution, le flou de mouvement (flou de bougé pour le mouvement de la caméra ou flou cinétique pour le mouvement du sujet), les occlusions, les variations de pose, les variations de luminosité et d’ombres, et enfin les variations d’angle d’inclinaison. Par ailleurs, quelques travaux se sont penchés spécifiquement sur le cas de la ré-identification par imagerie aérienne [15, 9, 33, 2, 26, 24], certains proposant en même temps des jeux de données dédiés (cf. Section 6.1.2).

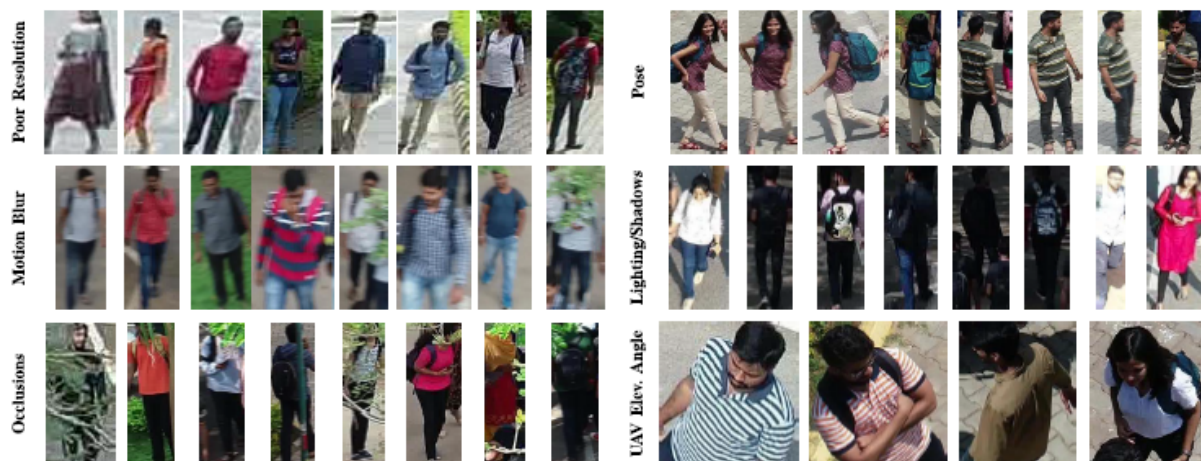


FIGURE 6.7 – Illustration des principaux défis de la ré-identification de personnes avec des exemples issus du jeu de données P-DESTRE (d’après [2]).

## 6.2 Décomposition en parties et comparaison par appariement des parties

Nous proposons ici une solution au problème de ré-identification qui se base sur l’appariement des parties des objets en fonction de plusieurs caractéristiques, incluant notamment la configuration spatiale. Nous nous focalisons en particulier sur le cas de la ré-identification de personnes qui intéresse l’entreprise Magellium, et utilisons pour cela une solution par décomposition en parties du corps, avec une segmentation basée sur la couleur. Après une présentation de notre approche, nous proposons une expérimentation afin de valider l’intérêt de celle-ci, en l’évaluant sur des séquences d’images aériennes de la base *Aeroscapes*.

### 6.2.1 Approche proposée

Notre idée est d’utiliser une approche par appariement basé sur des descripteurs, en y intégrant des descripteurs dédiés à la configuration spatiale. Ce type de descripteurs est intéressant pour reconnaître des objets ou des scènes avec des configurations spatiales particulières, ou pour faciliter l’appariement des parties si on utilise une décomposition en parties d’objets rigides ou peu déformables. Les cas d’usages identifiés pour la ré-identification à l’aide de la configuration spatiale couvrent à la fois la reconnaissance de scènes en tant qu’ensemble d’objets ou de points, ou d’objets en tant qu’ensemble de parties ou de points d’intérêt (cf. Tableau 6.1 et Section 1.1.3.1).

#### Comparaison avec la solution existante

La solution initiale proposée par Magellium se décompose ainsi, à partir d’une détection d’objet : segmentation objet/fond, description de l’objet par différentes caractéristiques, comparaison à un prototype et, pour l’apprentissage, ajout à la base si la similarité avec le prototype actuel est suffisante. Une première version se basait sur des descripteurs colorimétriques et géométriques de l’objet, puis des descripteurs de relations spatiales entre les parties colorimétriques de l’objet ont été ajoutés, mais sans avoir pu aboutir à des résultats satisfaisants. L’objectif fixé était d’améliorer la description des relations spatiales entre parties, en gardant si possible la même chaîne globale que la solution Magellium initiale. Cependant, une étude de cette solution a montré qu’il n’était pas suffisant de modifier uniquement cette description, mais qu’il était nécessaire avant cela de modifier la méthode de décomposition en parties, cette dernière impactant directement la description. Par ailleurs, nous avons aussi proposé

une amélioration de l'étape d'appariement intervenant pour la comparaison des configurations de deux acquisitions, en utilisant une combinaison de plusieurs caractéristiques des parties, plutôt que leur position uniquement. Enfin, nous avons remplacé la description de la configuration spatiale existante par un descripteur de position relative, plus récent et plus simple à comparer (cf. Chapitre 3). Ces améliorations ont fait l'objet d'une étude expérimentale en *python*, avant d'être éventuellement intégrées dans la solution Magellium en *C++*.

### Décomposition en parties

Notre approche se base sur une décomposition de la scène ou de l'objet en parties, afin de les décrire et de réaliser une comparaison par appariement de ces parties. L'approche pour la décomposition n'est alors pas la même dans le cas d'une scène ou d'un objet :

- décomposer une scène consiste à extraire des objets ou des régions, ce qui peut être réalisé par une méthode de segmentation sémantique classique, par exemple un modèle *Mask R-CNN* pré-entraîné sur des images similaires ;
- décomposer un objet consiste à extraire des parties significatives de l'objet, ce qui nécessite un modèle adapté à cet objet ou d'utiliser des caractéristiques permettant de distinguer ces parties.

Dans le cas de la ré-identification de personnes, qui sont assimilées à des objets animés, nous avons dû faire face au problème de la décomposition en parties, insuffisante dans la solution existante basée sur la couleur. Ce problème courant a été largement étudié depuis 20 ans, et des solutions récentes permettent d'obtenir une décomposition fine des parties du corps ou du squelette, en y associant l'estimation de pose typiquement [6, 28, 21, 16] (voir *BodyPix*<sup>1</sup> par exemple). Cependant, ces solutions sont conçues pour des images acquises dans un environnement "simple" et non pour de l'imagerie aérienne, et nécessitent notamment des résolutions supérieures. Nous sommes donc restés sur une approche basée sur les couleurs, mais en utilisant une méthode de classification par *clustering* dans l'espace de couleurs HSV, plutôt que la quantification fixe sur la teinte qui était utilisée. Cette approche est décrite dans le Chapitre 5. Notre solution donne des résultats visuellement satisfaisants sur le jeu de données *Aeroscapes*, ce qui a permis de générer un découpage en parties utilisable pour calculer différentes caractéristiques et évaluer l'appariement sur ces caractéristiques.

### Comparaison basée sur les parties

Afin de comparer deux acquisitions, la solution initiale utilisait une méthode spécifique pour chaque type de descripteur, sur l'objet entier pour les descripteurs de couleur et de forme, et par relation entre parties pour le descripteur de configuration. Pour ce dernier, elle passait par une étape d'appariement des parties en fonction de la position de leurs barycentres, puis la comparaison s'effectuait relation par relation. Nous proposons une solution utilisant les mêmes types de descripteurs mais entièrement basée sur l'appariement des parties, en calculant ces descripteurs sur les parties et en considérant simultanément les différents types de descripteurs lors de l'appariement. L'idée est ici de combiner plusieurs types de preuves pour établir la correspondance, de façon à éviter les erreurs, en utilisant pour cela les caractéristiques qui ne varient pas ou qui varient peu. La comparaison finale peut alors se faire en exploitant l'ensemble des caractéristiques ou juste celles qui varient. Par ailleurs, cette méthode peut aussi être utilisée en complément de descripteurs sur l'ensemble de l'objet comme des descripteurs de forme, de façon à garder une modélisation de la silhouette.

Le choix des descripteurs pour l'appariement et pour le calcul du score final dépend donc de ce qu'on attend comme variations et comme invariants dans les données. Afin de l'aiguiller, nous proposons dans le Tableau 6.1 une estimation de la fréquence et de l'importance de ces variations selon le cas d'usage. Suivant cette estimation, pour la ré-identification de personnes on peut se baser sur la caractéristique

1. <https://github.com/tensorflow/tfjs-models/tree/master/body-segmentation>

de couleur pour l'appariement, et éventuellement sur celle de configuration spatiale, ses variations dépendant de la résolution.

	objet rigide (ex : véhicule)		objet déformable (ex : personne)		scène (ex : complexe, convoi)	
	suivi	ré-identification	suivi	ré-identification	suivi	ré-identification
couleur des parties	rare/faible	rare/faible	rare/faible	(rare)/faible	rare/faible	rare/faible
forme des parties	assez rare et faible	assez rare et faible	assez fréquent et important	assez fréquent et important	assez rare et faible	assez rare et faible
agencement des parties	assez rare et faible	fréquent mais modéré	fréquent mais modéré	fréquent mais modéré	assez rare et faible	variable

TABLEAU 6.1 – Estimation de la fréquence et de l'importance des variations des caractéristiques pour le suivi et la ré-identification, selon le type d'objet. Pour une personne, la variation de la couleur est faible sauf dans le cas où elle changerait de tenue, cas qui concerne la ré-identification long terme.

Pour l'appariement en lui-même, l'utilisation de descripteurs par parties permet de calculer des distances entre les parties de deux acquisitions. En calculant ces distances pour chaque paire (au nombre de  $n \times m$  avec  $n$  et  $m$  le nombre de parties de chaque objet), on se ramène à un problème d'affectation linéaire (cf. Figure 3.3b), qui peut être résolu en temps polynomial (en  $O((n + m)^3)$ ) avec l'algorithme hongrois [14, 19] malgré un nombre de possibilités égal à  $\max(n, m)!$ . Dans la solution de Magellium, l'appariement est effectué en fonction de la position des barycentres des parties uniquement, avec une approche gloutonne. Une solution similaire est proposée dans [5], qui réalise un appariement en fonction de la forme (cf. Figure 3.4b), après un pré-appariement en fonction de la couleur. Nous proposons quant à nous de combiner les différents descripteurs retenus dans une même distance afin de tenir compte de chacun d'entre eux simultanément. Pour cela, il est nécessaire d'adapter les descripteurs à l'utilisation d'une distance globale, par linéarisation et normalisation typiquement, ou de définir une distance adaptée pour chaque type de descripteur afin d'en faire la somme. Il convient alors de fixer les coefficients de chaque critère en fonction de l'impact que l'on souhaite lui donner.

### Description de la configuration spatiale

Afin de décrire la configuration spatiale entre les parties, la solution Magellium était basée sur une quantification des relations topologiques d'Allen entre deux objets, en utilisant pour cela les distances entre les projections des objets selon les deux axes de coordonnées, s'inspirant de la méthode des 2D-strings [3]. La comparaison de deux configurations passait alors par des formules complexes définies pour chaque couple de relations topologiques possible ( $13 \times 13$  possibilités pour chacun des deux axes), ce qui était assez lourd. Nous suggérons quant à nous d'utiliser l'histogramme de forces, qui est un descripteur quantitatif plus simple et plus complet (cf. Section 2.1.1.2), avec l'approche "un contre tous" introduite dans la Section 3.2.1. Cela permet d'obtenir un descripteur par partie plutôt que par paire, ce qui réduit la complexité et permet de rester dans le cadre d'un problème d'affectation linéaire pour l'appariement. Nous comparons aussi cette solution à une approche barycentrique, qui pourrait être suffisante selon le cas d'usage.

### Apprentissage incrémental de prototypes

Enfin, la dernière partie de l'algorithme de ré-identification concerne l'apprentissage de prototypes de façon incrémentale. L'objectif est de pouvoir ajouter des éléments au fur et à mesure que ceux-ci sont découverts lors du traitement d'une vidéo, en identifiant s'ils ont déjà été vus ou pas, et en les ajoutant au modèle de l'objet correspondant si c'est le cas. Pour cela, la solution Magellium utilise le score de comparaison pour filtrer les nouveaux éléments selon ceux qui sont proches d'un prototype



existant ou pas. Si l'élément est suffisamment proche d'un prototype, le modèle est mis à jour avec le nouvel élément, sinon un nouveau prototype est initialisé à partir de l'élément. Cette mise à jour se fait par un algorithme de *clustering* par *k*-moyennes sur les descripteurs de l'ensemble des éléments, à chaque ajout d'un nouvel élément.

En gardant la même approche globale basée sur des prototypes, nous proposons quant à nous d'utiliser simplement le centroïde des descripteurs des éléments de la classe plutôt que de passer par un *clustering* à chaque mise à jour. L'ajout d'un élément peut alors se faire sans garder de trace des autres éléments de la classe, en conservant uniquement leur centroïde (le prototype) et leur nombre, selon l'équation 6.1. À noter qu'il est nécessaire de normaliser les caractéristiques si elles ne le sont pas encore (par exemple pour les histogramme de forces) avant de réaliser cette moyenne.

$$new\_prototype = \frac{n}{n+1} prototype + \frac{1}{n+1} new\_item \quad (6.1)$$

## 6.2.2 Expérimentations et résultats

### 6.2.2.1 Données et protocole expérimental

Dans notre évaluation, nous avons utilisé le jeu de données *Aeroscapes*<sup>2</sup> [20]. Ces données sont présentées dans la Section 1.2.2.3 et des extraits de ces séquences sont donnés dans la Figure 1.11. Les personnes sont extraites grâce aux masques de segmentation associés, puis décomposées en parties selon l'approche présentée dans le Chapitre 5. Nous avons en particulier utilisé trois séquences où l'on peut suivre une personne, différente à chaque fois. Des extraits sont donnés dans les Figures 1.12 et 1.13.

Notre expérimentation ayant seulement vocation à valider la faisabilité de l'approche, nous nous sommes contentés de mesures simples. Ainsi, nous avons utilisé les descripteurs suivants :

- un descripteur basé sur la couleur, en utilisant les valeurs HSV moyennes de la partie,
- un descripteur de forme basique, en prenant l'aire de la partie,
- des descripteurs de configuration spatiale : la position du barycentre de la partie dans un premier temps (donc sans considérer les autres parties), et l'histogramme de forces entre la partie et l'ensemble des autres dans un second temps (cf. Section 3.2.1).

Les valeurs de ces descripteurs pour deux images de test sont données dans la Figure 6.8, tandis que leurs histogrammes de forces "un contre tous" sont donnés dans la Figure 3.7. Ceux-ci sont ensuite normalisés selon la taille de l'image pour pouvoir être comparés.



	part	area	x_c	y_c	H	S	V	part	area	x_c	y_c	H	S	V	
	1	498	35	16	113	117	62	1	527	46	11	115	107	69	
	2	141	25	23	168	138	183	2	2021	46	45	102	59	237	
	3	1427	29	44	103	60	232	3	222	42	27	168	151	201	
	4	960	33	78	105	119	187	4	1092	51	87	105	120	197	
	5	97	29	110	95	44	246	5	184	52	133	95	39	245	

FIGURE 6.8 – Exemples de descriptions d'objets par les caractéristiques des parties, sur deux images extraites du jeu de données *Aeroscapes* et décomposées en parties. Les caractéristiques sont dans l'ordre : l'aire de la partie, les coordonnées ( $x_c$  et  $y_c$ ) de son barycentre et les valeurs HSV moyennes de ses pixels. Les parties à appairer sont identifiées avec la même couleur dans les deux tableaux.

Afin de valider notre approche, nous avons mené plusieurs tests de comparaison d'objets (personnes) et d'apprentissage de prototypes, sur plusieurs séquences d'images. Pour la comparaison, nous

2. <https://github.com/ishann/AeroScapes>

avons comparé les scores d'appariement (distances totales obtenues avec l'algorithme hongrois) pour quelques images par rapport à la même référence, en utilisant les différents descripteurs séparément ou combinés (avec une pondération arbitraire des critères). Ces images sont présentées dans la Figure 6.9. Pour l'apprentissage, nous avons utilisé toute une séquence contenant la même personne pour calculer un prototype, puis mesuré l'évolution du score de comparaison de chaque image avec ce prototype, de manière à identifier les éléments les plus proches et les plus éloignés. Pour terminer, nous avons comparé ce prototype à des images d'une autre classe (une autre séquence) pour vérifier que la méthode était capable de les distinguer.

### 6.2.2.2 Résultats et discussion

Les résultats des comparaisons aux prototypes des séquences 040000 et 000002 sont donnés dans les Tableaux 6.2 et 6.3 respectivement, pour les images de la Figure 6.9 et pour différents critères. Ceux-ci représentent les distances totales d'appariement, qui ont été normalisées selon la plus grande distance atteignable pour chaque critère, multipliée par le nombre de parties appariées. Cela dit, la comparaison des valeurs entre différents critères n'est pas forcément pertinente, du fait de leur signification différente. En revanche la comparaison des distances pour un critère donné (par ligne) a un sens, ce qui permet d'ordonner les éléments en fonction de leur similarité au prototype. À noter également que le score ne reflète pas forcément l'appariement attendu, puisqu'il correspond au score du meilleur appariement obtenu, qui peut être différent selon le critère utilisé.

Pour le prototype1 obtenu sur la séquence 040000 (Tableau 6.2), les images "img1" et "img2" issues de cette séquence sont très proches pour l'ensemble des critères (sauf pour la position des barycentres pour l'image "img2", où les parties se sont décalées sur la droite), ce qui était attendu. L'image "img3" est un peu plus éloignée mais reste assez proche quelque soit le critère, ce qui n'est pas surprenant non plus étant donné qu'elle est aussi assez proche visuellement. En revanche, l'image "img4" est plus éloignée en ce qui concerne les caractéristiques d'aire et de position, ce qui est normal étant donné qu'une bonne partie de l'objet est masquée. Quant à l'image "img5" issue de l'autre séquence, elle est éloignée en termes de couleur et également d'aire, mais pas vraiment en termes de configuration (position des barycentres et histogrammes de forces), ce qui est dû à l'allure similaire de la personne dans cette image par rapport aux images de la séquence 040000. Enfin, l'appariement avec le prototype2 se montre plus complexe, avec des distances élevées pour l'ensemble des critères sauf pour l'histogramme de forces, ce qui montre que la configuration est bien similaire. La distance plus élevée ici pour la position des barycentres est sans doute due au fait que davantage de parties sont présentes sur le reste de la séquence 000002, notamment du fond qui n'a pas été filtré, ce qui perturbe moins l'histogramme de forces "un contre tous". Pour le prototype2 (Tableau 6.3), l'image "img5" est correctement appariée et toutes les images de la séquence 040000 sont éloignées (en particulier l'image "img4") quelque soit le critère, ce qui est cohérent. Cela dit, les images "img1", "img2" et "img3" sont un peu moins éloignées en termes de configuration, ce qui pourrait permettre d'envisager un appariement selon ce critère.

De façon générale, l'appariement en fonction de la couleur donne des distances faibles pour les images de la même séquence, et éloignées pour des images de l'autre séquence, ce qui laisse penser que ce critère seul pourrait permettre de distinguer des objets. Cependant, les couleurs pourraient être présentes mais associées à d'autres parties, ce qui a conduit à utiliser d'autres critères (*i.e.*, l'approche "sacs de caractéristiques" indépendantes ne suffit pas). L'utilisation de caractéristiques comme l'aire et la position des parties permet de pallier cela, mais peut alors empêcher de reconnaître un même objet vu dans des configurations différentes. On vérifie donc l'importance du choix des caractéristiques et de leur pondération dans la décision finale. En particulier, pour ce cas d'usage où le mouvement des parties peut être important, il n'est pas utile de considérer la forme et la configuration précise des parties, la position des barycentres étant suffisante si l'objet est visible en entier. Pour aller plus loin, il serait alors utile de chercher à détecter les occultations, en amont ou en même temps que l'appariement.

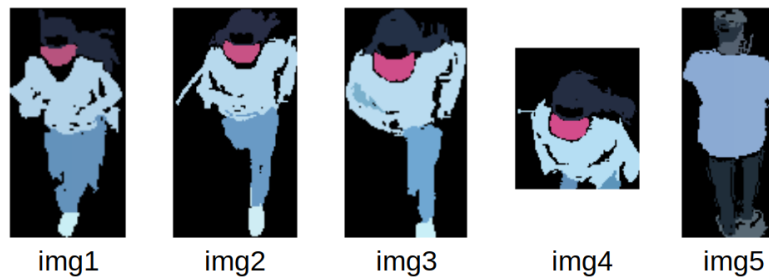


FIGURE 6.9 – Images de test extraites des séquences 040000 et 000002 du jeu de données *Aeroscapes*. Les images "img1", "img2", "img4" contiennent 5 parties, et les images "img3", "img5" en contiennent 6.

critères	img1	img2	img3	img4	img5	prototype2
aire	0.28	0.28	0.32	0.44	0.38	0.42
couleur	0.11	0.18	0.23	0.24	0.47	0.51
position	0.11	0.27	0.21	0.91	0.23	0.67
F-histo	1.43	1.46	2.00	2.45	1.75	1.59
toutes	0.24	0.30	0.40	0.76	0.51	0.60

TABLEAU 6.2 – Distance entre le prototype1 et plusieurs images de test avec différents critères. En rouge les distances élevées, en orange les distances moyennement élevées.

critères	img1	img2	img3	img4	img5	prototype1
aire	0.40	0.41	0.44	0.51	0.12	0.42
couleur	0.40	0.41	0.44	0.52	0.09	0.51
position	0.27	0.31	0.21	0.43	0.13	0.67
F-histo	1.99	1.98	2.17	2.36	1.68	1.59
toutes	0.54	0.53	0.60	0.83	0.26	0.60

TABLEAU 6.3 – Distance entre le prototype2 et plusieurs images de test avec différents critères. En rouge les distances élevées, en orange les distances moyennement élevées.

### 6.3 Conclusion

La ré-identification est un problème assez récent en vision par ordinateur, mais finalement très proche du problème général de reconnaissance à grain fin à partir de peu de données, auquel elle ajoute la problématique de robustesse aux conditions d'acquisition. La solution proposée par Magellium adopte une approche classique en reconnaissance d'objets qui consiste à décrire et comparer différentes caractéristiques de ces objets, en ajoutant cependant des caractéristiques dédiées à la configuration des parties. L'idée est d'obtenir une certaine robustesse aux variations de point de vue en considérant des caractéristiques moins variables comme la forme globale (silhouette) ou les couleurs présentes. En se basant sur cette solution, nous avons proposé plusieurs modifications, notamment dans le but d'améliorer la prise en compte de la configuration spatiale.

La première modification a concerné la décomposition en parties, pour laquelle nous avons proposé une nouvelle solution plus performante et plus robuste aux variations, en utilisant une approche par *clustering* dans l'espace HSV conique (cf. Chapitre 5), plutôt qu'un seuillage sur la teinte. La deuxième modification a concerné la méthode de comparaison elle-même : au lieu de calculer des descripteurs globaux et de les utiliser séparément, nous avons proposé de les calculer sur les parties et de les utiliser conjointement pour appairier celles-ci, de façon à mieux exploiter la décomposition en parties et améliorer la robustesse de l'appariement. Nous retrouvons alors un problème d'affectation linéaire, pouvant être résolu avec l'algorithme hongrois, ce qui permet d'obtenir directement un score de similarité (ou



une distance). La troisième modification a concerné la description de la configuration spatiale : pour remplacer la description trop complexe, nous avons proposé d'utiliser des histogrammes de forces, avec l'approche "un contre tous" introduite dans le Chapitre 3, c'est-à-dire en calculant pour chaque partie l'histogramme entre celle-ci et l'ensemble des autres. Enfin, l'apprentissage de prototypes est réalisé par une simple moyenne des descripteurs, plutôt qu'en faisant appel à un algorithme de *clustering* à chaque ajout.

Nous avons évalué notre solution sur le cas d'usage de la ré-identification de personnes par imagerie aérienne qui intéresse Magellium, en utilisant des séquences d'images du jeu de données *Aeroscapes*. Les premiers résultats ont pu montrer l'intérêt de l'approche, qui doit maintenant être évaluée sur davantage de données. Par ailleurs, cette approche a l'avantage d'être explicable et de ne pas demander d'importantes ressources de calcul (excepté pour le calcul des histogrammes de force, selon la précision choisie). Cependant, elle n'est pas forcément la plus adaptée pour la ré-identification de personnes, où les différences intra-classe (pour une même personne) peuvent être élevées, tandis que les différences inter-classe sont relativement faibles. D'une manière générale, le problème de la ré-identification cherchant à reconnaître des exemples très particuliers d'objets ou de groupes d'objets, les solutions basées sur la reconnaissance fine ne suffisent pas. Une autre approche possible et que nous jugeons plus pertinente serait de rechercher des détails caractéristiques, c'est-à-dire qui apparaissent rarement dans les données, avec une analyse sur des données similaires ou en introduisant de la connaissance métier. Cependant, ce type de détail n'existe pas toujours suivant la résolution des images.



# Conclusion



# Conclusion

*Cela est bien, mais il faut cultiver notre jardin.*

*Candide, Voltaire*

Cette thèse a abordé les problèmes de la description de configurations spatiales et de leur exploitation dans des chaînes de traitement de reconnaissance de scènes ou d'objets. La configuration spatiale des objets d'une scène ou des parties d'un objet est un aspect particulièrement important dans la compréhension de celle-ci, avec un champ d'applications important. Or, ce sujet est encore peu abordé à ce jour, y compris par les approches les plus modernes comme les CNN. Deux approches principales sont alors possibles pour prendre en compte la configuration spatiale dans les chaînes de traitement : entraîner les CNN sur des données adaptées, c'est-à-dire contenant les configurations attendues, ou revenir à des approches "traditionnelles" utilisant des descripteurs définis "manuellement", en ajoutant des descripteurs dédiés à la configuration.

La première approche a l'avantage d'être entièrement automatique, mais les défauts d'être peu explicable et de nécessiter de grandes quantités de données d'apprentissage pour généraliser correctement, ce qui peut poser problème pour certaines applications. À l'inverse, la seconde approche est plus générique, naturellement explicable en considérant des caractéristiques bien définies, et elle nécessite moins de données pour la classification. Nous nous sommes donc intéressés à celle-ci, en explorant l'utilisation de descripteurs de configuration spatiale. Cette approche nous a également conduits à traiter le problème de la segmentation, nécessaire pour décrire précisément ces configurations. Enfin, nous avons eu l'occasion d'évaluer nos solutions sur une application concrète issue du contexte CIFRE de cette thèse : la ré-identification de personnes, afin d'améliorer une solution existante.

## Bilan global

Le sujet principal de cette thèse a été d'explorer des modélisations de la configuration spatiale des scènes ou des objets. Les descripteurs dédiés à la configuration spatiale sont étudiés depuis plusieurs décennies, mais ils sont encore très peu utilisés dans les applications. Le Chapitre 1 de ce manuscrit introduit le sujet de la configuration spatiale, ainsi que ses usages possibles et les défis auxquels il faut faire face pour l'utiliser dans des tâches de compréhension de scènes. Ces défis sont notamment le fait que ces descripteurs ne traitent que des configurations de deux objets et le fait qu'ils ne peuvent être calculés que sur des formes binaires. Dans le but d'exploiter la configuration spatiale pour diverses applications, cette thèse a eu pour objet principal d'explorer des approches permettant de dépasser ces limitations. Nous revenons sur l'ensemble des défis dans le paragraphe suivant.

Dans un premier temps, nous avons commencé par considérer les descripteurs de relations spatiales entre deux objets (cf. Chapitre 2), pour lesquels l'historique est le plus important, avec par exemple l'histogramme de forces introduit en 1999, qui permet de décrire précisément la configuration avec une approche directionnelle. Les approches les plus récentes se basent maintenant sur la fusion de descriptions directionnelles et topologiques, comme le  $\phi$ -descripteur qui est une extension de l'histogramme de forces. Nous avons quant à nous proposé une autre extension de ce dernier dédiée aux configura-

tions complexes : le bandeau de forces. Ce descripteur peut alors être utilisé pour évaluer des relations spatiales en langage naturel, ce que nous avons vérifié expérimentalement, ou dans une chaîne de traitement plus complexe pour y intégrer la configuration spatiale. Ces travaux ont fait l'objet de publications dans une conférence et un journal internationaux [6, 9], ainsi qu'une conférence nationale [5].

Nous avons ensuite abordé la problématique du passage à des modélisations de scènes comportant plusieurs objets (cf. Chapitre 3). Quelques solutions existent pour cela, comme les sacs de relations ou les graphes relationnels attribués (ARG), mais elles n'ont pas encore reçu énormément d'attention et peuvent encore être améliorées. En particulier, les représentations par graphes se heurtent au problème de l'appariement, qui peut devenir très complexe lorsque le nombre d'éléments est important. Nous avons proposé une solution simplifiant ce problème en utilisant des descripteurs de relations "un contre tous", c'est-à-dire d'une partie par rapport à l'ensemble des autres. Cette solution permet de se ramener à un problème d'affectation linéaire, qui peut alors être résolu en temps polynomial avec l'algorithme hongrois. Des tests préliminaires utilisant l'histogramme de forces ont montré l'intérêt de notre approche, qui a aussi l'avantage de pouvoir se combiner avec d'autres types de descripteurs.

Par ailleurs, les descripteurs de configuration spatiale sont en général calculés sur des formes binaires, ce qui nécessite une étape de segmentation en amont. Deux cas se présentent alors : celui d'une scène dont on souhaite extraire les objets et celui d'un objet qu'on souhaite décomposer en parties. En général, le premier cas peut être traité facilement par un des nombreux modèles de segmentation pré-entraînés qui sont maintenant accessibles, pour des scènes variées contenant toutes sortes d'objets. Le deuxième cas est plus complexe car en général il n'existe pas de modèle dédié, excepté pour quelques applications comme la décomposition en parties du corps. Le cadre industriel de cette thèse nous a amené à traiter ce cas, mais sur des données peu résolues où ces modèles n'étaient pas utilisables. Nous avons donc dû utiliser une autre approche, et avons développé pour cela une solution basée sur la couleur, grâce à un algorithme de *clustering* dans l'espace HSV conique, introduisant ici la notion d'espace de mesure adapté (cf. Chapitre 5). Par ailleurs, nous avons aussi dû aborder le problème de la segmentation objet-fond afin d'utiliser des données contenant déjà des annotations de détection sous forme de rectangles englobants. Nous avons proposé pour cela une solution combinant une segmentation sur toute l'image et une sélection du meilleur segment, en se basant sur plusieurs critères géométriques et sémantiques (cf. Chapitre 4). Cette approche et son extension ont fait l'objet de publications dans deux conférences internationales [7, 17] et un atelier d'une conférence nationale [8].

Pour finir, le contexte industriel de cette thèse nous a conduits à traiter une application au problème de la ré-identification d'objets (cf. Chapitre 6), afin d'améliorer une solution développée auparavant et utilisant plusieurs types de descripteurs des objets à reconnaître, dont des descripteurs dédiés à la configuration spatiale de leurs parties. Après une étude approfondie de la solution actuelle, nous avons proposé une nouvelle solution entièrement basée sur l'appariement des parties des objets entre deux acquisitions, en l'associant à une décomposition de la scène en parties et en calculant les descripteurs sur ces parties. A nouveau, l'appariement est réalisé grâce à l'algorithme hongrois, simultanément sur les différents types de descripteurs en combinant les distances pour chaque type. Il est alors possible d'utiliser les descripteurs définis précédemment afin de mieux prendre en compte la configuration spatiale. Nos évaluations de cette solution constituent une preuve de concept de l'utilisation de tels descripteurs pour une tâche de ré-identification.

## Retour sur les défis de la configuration spatiale

Comme nous avons pu le voir, la configuration spatiale est un aspect important pour la compréhension de scènes, avec un vaste champ d'étude et d'applications allant de la simple description de scène à la navigation autonome, en passant par la reconnaissance d'objets ou la recherche dans une base. Leur modélisation a été largement étudiée depuis ses débuts il y a un demi-siècle avec la première catégorisation des relations par Freeman. Ainsi, de nombreux modèles ont été proposés pour décrire les relations

topo-directionnelles entre deux objets, avec des solutions complètes et efficaces comme l'histogramme de forces ou le  $\phi$ -descripteur. Plusieurs modèles ont aussi été introduits pour décrire des configurations spatiales de scènes comportant plusieurs objets, comme les graphes relationnels attribués ou les sacs de relations. En revanche, l'utilisation de tous ces modèles pour les tâches de compréhension de scènes reste encore peu exploré, malgré un intérêt certain. Cela est dû à la concurrence d'autres approches performantes comme l'apprentissage profond sur de grandes bases de données, mais aussi à plusieurs limitations inhérentes à ces modèles.

Plusieurs défis se posent alors et ont été résumés dans la Section 1.1.5. Nous revenons ici sur chacun d'eux à l'aune de nos travaux, en mettant en avant nos contributions, leurs limitations et perspectives :

**1. la modélisation de relations et de configurations spatiales pour des configurations complexes ou ambiguës :**

Dans le Chapitre 2, nous avons proposé un descripteur dédié aux configurations complexes entre deux objets avec le bandeau de forces, qui est une extension de l'histogramme de forces combinant tout un panel de paramètres de forces dans un descripteur 2D. Étant donné que des forces différentes décrivent des interactions différentes entre les objets, ce descripteur permet d'avoir une description plus complète, capable de traduire des configurations complexes. Nos évaluations sur la classification de relations spatiales directionnelles simples ont permis de montrer la possibilité d'exploiter ce descripteur avec un réseau convolutionnel (CNN). Cependant, elles restent à compléter par un autre type d'évaluation et avec des configurations complexes plus variées pour mieux évaluer la performance de celui-ci. Ce type de descripteurs peut être étendu à des configurations spatiales entre plusieurs objets, selon l'approche détaillée dans le point suivant.

**2. le passage de descripteurs de relations spatiales entre deux objets à des descriptions de configurations de scènes comportant plusieurs objets :**

Dans le Chapitre 3, nous avons proposé une approche de description de la configuration spatiale par parties, en considérant la relation de chaque partie de la scène face à l'ensemble des autres. Cette approche "un contre tous" permet d'obtenir une description similaire à celles utilisées pour les caractéristiques courantes, et de comparer deux scènes en comparant leurs descripteurs. Nous suggérons alors de comparer deux scènes par appariement des parties en exploitant ces descripteurs, en utilisant l'algorithme hongrois pour résoudre ce problème d'affectation linéaire. Cette méthode a été évaluée sur quelques images en utilisant l'histogramme de forces comme descripteur de relations, confirmant l'intérêt de l'approche, mais reste à évaluer plus précisément sur davantage de données, avec des configurations variées.

**3. l'utilisation de descripteurs de configuration spatiale sur des images réelles non segmentées :**

Afin d'utiliser les descripteurs de relations spatiales existants pour les images réelles, la solution classique consiste à les segmenter. Dans nos travaux, nous proposons deux approches adaptées à deux cas différents : celui de scènes dont on dispose déjà d'annotations de détection d'objets (rectangle englobant et/ou classe) et celui d'objets segmentés que l'on souhaite décomposer en parties. Dans le premier cas, qui consiste en une segmentation objet-fond, nous avons proposé d'utiliser un modèle de segmentation existant sur l'image entière puis d'extraire l'élément le plus pertinent à l'aide des annotations de détection, en utilisant une combinaison de critères géométriques et sémantiques. Cette solution, présentée dans le Chapitre 4, offre une alternative à la méthode GrabCut dans le cas où l'on dispose d'un modèle de segmentation pré-entraîné adapté. Pour le second cas, des modèles existent mais dépendent du type d'objet donc ne sont pas adaptés à tous les cas d'usage. Souhaitant avoir une approche utilisable pour n'importe quel type d'objets, nous avons proposé une méthode de segmentation non supervisée basée sur la couleur, notamment sur la composante de teinte présente dans les espaces de couleur polaires de type



HSV. Ces approches générales de segmentation offrent une solution pour utiliser des images réelles pour le calcul de descripteurs de relations spatiales.

Par ailleurs, concernant les approches de bout-en-bout visant à décrire les configurations directement depuis des images non segmentées, nous avons pu constater quelques tentatives récentes dans la littérature avec la détection de triplets (sujet, relation, objet) mais leurs performances ne sont pas encore satisfaisantes. Le développement récent de jeux de données importants comme *Rel3D* pourrait néanmoins permettre d'aller vers ce type de solutions.

**4. la prise en compte de la pose des objets et du point de vue dans la scène 2D, avec possibilité d'invariance selon la tâche :**

Dans nos travaux, nous avons mis en avant l'existence de descripteurs de relations spatiales robustes aux variations de point de vue, notamment l'histogramme de forces qui est naturellement invariant aux translations et peut être normalisé pour être rendu invariant à toutes les similitudes. Dans le Chapitre 2, nous avons introduit le bandeau de forces, qui est destiné aux configurations complexes et bénéficie des mêmes propriétés, bien que nous n'ayons pas vérifié cela dans nos expérimentations. En revanche, nous avons pu montrer leur capacité à distinguer des points de vue différents lorsque l'on utilisait pas de normalisation, en classifiant des configurations selon leur direction. Ces descripteurs étant destinés aux relations entre deux objets, leur utilisation dans des modèles de scènes comportant davantage d'objets suppose alors d'utiliser les mêmes paramètres de normalisation pour toutes les relations. L'utilisation de tels modèles n'a cependant pas encore été évaluée, que ce soit dans nos travaux ou dans la littérature. Il serait donc intéressant d'explorer cela et de montrer sur un cas concret la possibilité d'utiliser de tels modèles.

Par ailleurs, la prise en compte de la pose des objets dans une scène 2D est un sujet qui reste à explorer, de la même façon que dans une scène 3D.

**5. la prise en compte de la pose des objets et du point de vue dans la scène 3D, avec possibilité d'invariance selon la tâche :**

L'étude de modèles spécifiques aux scènes 3D n'a pas été abordée dans nos travaux, celles-ci étant traitées comme des scènes 2D, ce qui peut générer des erreurs dues aux déformations des objets selon le point de vue. Et bien que nous ayons mené quelques expérimentations sur des données 3D, celles-ci ne contenaient pas de variation de point de vue suffisant pour évaluer nos approches, si ce n'est pour la translation où l'invariance est facilement vérifiée.

Par ailleurs, la prise en compte de la pose des objets dans les configurations spatiales est un sujet qui reste à explorer, même si plusieurs jeux de données intégrant cet aspect ont été proposés récemment, comme *SpatialSense* ou *Rel3D*. Nos évaluations et celles de la littérature sur le jeu de données *SpatialSense* ont d'ailleurs montré l'importance de modéliser cet aspect pour décrire correctement la relation dans l'espace 3D. En parallèle, l'estimation de pose est un sujet qui a largement été étudié dans la littérature. Son intégration dans des modèles de configuration spatiale devrait donc être envisageable dans un futur proche.

**6. l'intégration des descripteurs dans des chaînes de traitement ou des modèles existants :**

La chaîne de ré-identification que nous avons proposée dans le Chapitre 6 est un parfait exemple d'utilisation de descriptions de la configuration spatiale dans une chaîne de traitement, celle-ci répondant à un cas d'usage concret lié au contexte industriel de cette thèse. Nous avons ainsi montré qu'une telle description pouvait être intégrée parmi d'autres descriptions de la scène, afin d'améliorer sa compréhension et de permettre une comparaison avec d'autres en prenant en compte cet aspect. Cependant, notre évaluation a concerné la ré-identification de personnes, où la configuration spatiale précise n'a que peu d'intérêt. D'autres applications pourraient alors être envisagés, comme la reconnaissance de paysages aériens pour la localisation basée vision.

Un autre exemple plus restreint concerne la description des relations spatiales entre objets en langage naturel, que nous avons traitée comme évaluation du bandeau de forces dans le Cha-

pitre 2. Néanmoins, d'autres tâches similaires pourraient être envisagées également, comme la recherche de configurations similaires dans une base d'images, ou de triplets (sujet, relation, objet) en l'associant à une description des objets.

#### 7. la disponibilité de données pour entraîner et évaluer les approches :

Afin d'évaluer nos approches, nous avons été confrontés au problème général de la disponibilité de données segmentées et annotées avec des relations spatiales. Nous avons alors généré et annoté un jeu de données synthétiques de relations entre deux objets et segmenté deux jeux de données existants, pour des tâches différentes donc avec des solutions différentes, qui sont détaillées dans les Chapitres 4 et 5. Ils nous ont alors permis de mener à bien différents types de tests et pourraient profiter à d'autres évaluations similaires. Par ailleurs, le développement de grande bases de données comme *SpatialSense* ou *Rel3D* devrait permettre de mieux évaluer les approches à l'avenir, pour la reconnaissance de relations prenant en compte la pose des objets. Enfin, il serait également intéressant d'évaluer l'intérêt de la configuration spatiale dans d'autres cas d'usage plus larges comme la reconnaissance d'objets ou la recherche dans une base, où de nombreux jeux de données sont disponibles.

## Perspectives globales

Ainsi, la recherche sur la modélisation des configurations spatiales a encore plusieurs défis à relever afin de pouvoir utiliser cette composante pour des applications concrètes. En particulier, deux défis majeurs sont le passage à des modèles de scènes intégrant leur aspect 3D et la pose des objets. Ces défis pourraient être abordés en exploitant des connaissances a priori sur les objets identifiés, comme leurs formes ou leurs configurations usuelles, afin d'estimer leur pose, ou en associant plusieurs vues de la même scène, comme dans [19, 18] pour la détection d'objets multi-vues. Ces approches pourraient alors être combinées aux modèles que nous avons étudiés, comme perspective de cette thèse.

Une autre perspective importante pour cette thèse et ses applications est la prise en compte de la dimension temporelle, qui conduit à de nouvelles applications avec la modélisation de configurations spatio-temporelles. Elle pourrait également être exploitée dans les modélisations classiques pour les rendre plus fiables, en analysant l'évolution des éléments présents et de leurs caractéristiques pour en déduire des invariants, ce qui pourrait permettre de traiter la perspective précédente en apportant des points de vues différents. En particulier, le cas de la vidéo permet en général d'avoir de nombreux exemples des mêmes objets avec des variations progressives, ce qui permet de les suivre et d'en générer des modèles plus robustes à ces variations.

Dans ce cadre, le cas de la ré-identification est intéressant car il considère généralement de telles données, où l'on dispose de quelques exemples proches pour générer le modèle de la scène ou de l'objet observé, par apprentissage par exemple. Dans notre solution basée sur l'appariement de parties grâce à plusieurs caractéristiques, il serait alors intéressant d'utiliser des caractéristiques différentes pour l'appariement et la comparaison, ou de les pondérer différemment, en fonction des variations ou invariances observées dans la scène ou l'objet d'intérêt. Cet aspect reste donc à explorer afin d'obtenir une preuve de concept pour cette application. Cette solution serait également à évaluer sur davantage de données et sur des cas d'usage différents, adaptés à l'exploitation de la configuration spatiale notamment, comme la reconnaissance de bâtiments particuliers dans des images aériennes, ou de vêtements particuliers dans des images de vidéo-surveillance.

Enfin, d'autres perspectives concernent les autres applications ayant besoin ou pouvant profiter d'une modélisation de la configuration spatiale, qui sont listées dans la Section 1.1.3. Une application générale concerne la tâche de reconnaissance d'objets ou de scènes, ainsi que les tâches très proches de recherche dans une base (ou CBIR) ou de détection de changement. Celle-ci peut alors s'intégrer dans de nombreuses autres applications plus spécifiques, comme la reconnaissance de paysages aériens pour la localisation basée vision, la reconnaissance de cellules particulières pour la médecine, la détection

de situation à risque pour l'aide à la navigation, la reconnaissance de stratégie en sport collectif, etc. Toutes ces perspectives militent pour une meilleure prise en compte de la configuration spatiale dans les tâches de compréhension de scène et pour un investissement accru de la recherche en vision par ordinateur sur ce sujet.

# **Annexes**



# Annexes

## Annexe A Historique des espaces de couleurs

La perception de la couleur a fait l'objet de nombreuses études depuis très longtemps, en sciences optiques puis en neurosciences d'une part, et pour des applications artistiques puis industrielles d'autre part. En particulier, la mesure de la perception humaine de la couleur est devenue un domaine scientifique à part entière, appelé colorimétrie. En science informatique, elle amène deux défis majeurs : retranscrire en grandeurs numériques les grandeurs physiques mesurées par les capteurs, et utiliser des représentations de données adaptées pour retranscrire la perception humaine. Dans cette optique, plusieurs espaces de couleurs numériques ont été développés, comme les espaces HSV et HSL, en s'appuyant sur les travaux antérieurs de catégorisation des couleurs sous forme d'atlas, comme ceux de Munsell [12] ou Ostwald [11]. Ceux-ci permettent de séparer la couleur en une composante chromatique (la teinte notamment) et une composante de luminosité, ce qui est particulièrement utile pour des tâches de reconnaissance où la luminosité peut varier.

Les premiers modèles de couleurs sont apparus avec la théorisation de l'art et l'étude de la perception humaine des couleurs d'une part, initiées par Léonard de Vinci ou Johann Wolfgang Goethe par exemple, et en sciences optiques d'autre part, initiées par les travaux d'Isaac Newton au début du 18<sup>e</sup> siècle, les deux visions étant liées par la démarche expérimentale.

En théorie de l'art, la production de la couleur a été étudiée à la même époque par plusieurs peintres et créateurs textiles. Les connaissances sur les mélanges de couleurs se sont développées à partir de la Renaissance (avec Léonard de Vinci par exemple), qui a révélé l'existence de trois couleurs primaires (rouge, bleu, jaune) permettant de produire toutes les autres. Suite à ces découvertes, les premières roues (ou cercles) chromatiques apparaissent au début des années 1700 (comme celles de Claude Boutet), montrant une continuité dans la variation des couleurs perçues et la notion de couleurs complémentaires (voir la Figure A.1). Plus tard, dans le but de répertorier toutes les couleurs possibles avec une intensité variable, Gaspard Grégoire conçoit le premier atlas des couleurs dès le milieu des années 1780 et le publie dans les années 1810 [14], devenant l'un des premiers à décrire la couleur en termes de teinte, de chrominance et luminosité, ce qui est toujours utilisé aujourd'hui. Dans la lignée de ces travaux, Philipp Otto Runge développe le concept de sphère de couleur en 1807 [4], et Goethe propose sa propre roue chromatique en 1810 (voir la Figure A.1b).

En science optique, Newton a été le premier à mener des expériences approfondies sur le comportement de la lumière. Il a publié sa théorie des particules dans son recueil "Opticks" en 1704 [1], incluant sa célèbre analyse de la réfraction des couleurs. Il fut aussi le premier à présenter les couleurs sur un cercle (voir la Figure A.1a), initiant un questionnement majeur dans l'art et un tout nouveau domaine de recherche. Un siècle plus tard, au début des années 1800, Thomas Young y ajouta la théorie ondulatoire de la lumière et proposa la théorie additive des couleurs, émettant l'hypothèse de l'existence de trois récepteurs de couleur dans la perception humaine (consacrés au rouge, au vert et au bleu) [2], ce qui fut prouvé en 1859 par Hermann Von Helmholtz, en observant par micro-spectrophotométrie trois types de récepteurs de couleur dans la rétine : les "cônes" [10]. La théorie de Young a donné naissance à la colorimétrie au milieu du XIX<sup>e</sup> siècle, dont le but est de transcrire la perception humaine des couleurs en

lois mathématiques, à l'image des quatre lois de Grassmann sur la trichromie, publiées en 1853 dans sa "Théorie du mélange des couleurs" [6]. Celle-ci a ensuite été développée par James Clerk Maxwell [7], qui fut le premier à proposer des équations pour décrire la couleur à partir de ses composantes RGB.

Pendant, une autre conception a été proposée par Goethe, qui fut en fait le premier à étudier la perception psychologique de la couleur, aboutissant à sa "Théorie des couleurs" en 1810 [3], mais manquant de mesures physiques pour être appelée colorimétrie. Outre son opposition à la théorie de Newton sur la génération des couleurs, Goethe a observé l'existence de couleurs opposées (ou "couleurs réciproques"), permettant de présenter les couleurs comme une roue chromatique à six couleurs principales (voir la Figure A.1b). Cette théorie a été développée par Ewald Herring [8], qui a postulé que les trois couleurs de base étaient ensuite traitées par la vision humaine pour produire en fait 4 couleurs de base, chacune ayant un opposé (rouge vs. vert et bleu vs. jaune), de façon telle que le mélange de couleurs opposées ne peut être vu comme un arrangement d'elles-mêmes (i.e., rouge-vert ou bleu-jaune, voir la Figure A.1c). La théorie des couleurs opposées n'a été prouvée que récemment par les neurosciences, révélant que certaines cellules du cortex visuel codent effectivement les couleurs de manière antagoniste [13, 15]. La perception psychologique des couleurs humaines a également été étudiée par Michel-Eugène Chevreul, qui a développé la loi des contrastes simultanés de couleurs en 1839, montrant l'impact des couleurs dans la zone environnante.



(a) Cercle des couleurs de Newton (b) Roue des couleurs de Goethe (c) Roue des couleurs de Herring

FIGURE A.1 – Évolution des modèles de couleurs principales.

Tous ces travaux ont développé plusieurs concepts qui se sont imposés et sont encore applicables de nos jours : l'existence de trois couleurs fondamentales (liées aux trois types de cônes de la rétine humaine), qui peuvent être mélangées pour produire toutes les autres, la sensation de couleurs opposées, l'importance du contraste, etc. Ils ont donné naissance aux premières descriptions de la couleur en termes de teinte, de chrominance et de luminosité, d'abord qualitatives, dans la suite des représentations circulaires de Newton et Goethe par exemple, puis quantitatives avec la colorimétrie. Ce type de modélisation a alors été développé par des techno-industriels comme Gaspard Grégoire, Michel-Eugène Chevreul, Ogden Rood ou Wilhelm Ostwald, qui ont permis l'exploitation de ces connaissances dans les arts et l'industrie en produisant des atlas de couleurs [14, 5, 9, 12, 11].

Parmi ces différentes catégorisations, celle qui eut le plus d'impact fut sans doute celle d'Albert Henry Munsell (voir les Figures A.5 et A.2), proposée en 1905, qui n'essaya pas d'intégrer des considérations physiques ou artistiques mais uniquement perceptives, en utilisant des secteurs de couleur représentatifs de la vision humaine [12]. Ce modèle est encore utilisé pour de nombreuses applications artistiques et est le système officiel d'identification des couleurs aux États-Unis. De la même façon, la théorie de Herring sur les couleurs opposées a donné naissance au système NCS (voir la Figure A.3), qui est le système officiel d'identification des couleurs en Suède et dans plusieurs autres pays. Le système d'Ostwald [11] se base aussi sur cette théorie mais en l'adaptant pour mieux traduire l'additivité des couleurs, se rapprochant alors de celui de Munsell. Il est aussi plus proche des systèmes numériques actuels dans le sens où il a cherché à donner une représentation uniforme, qui prend la forme d'un double cône (voir la Figure A.4).



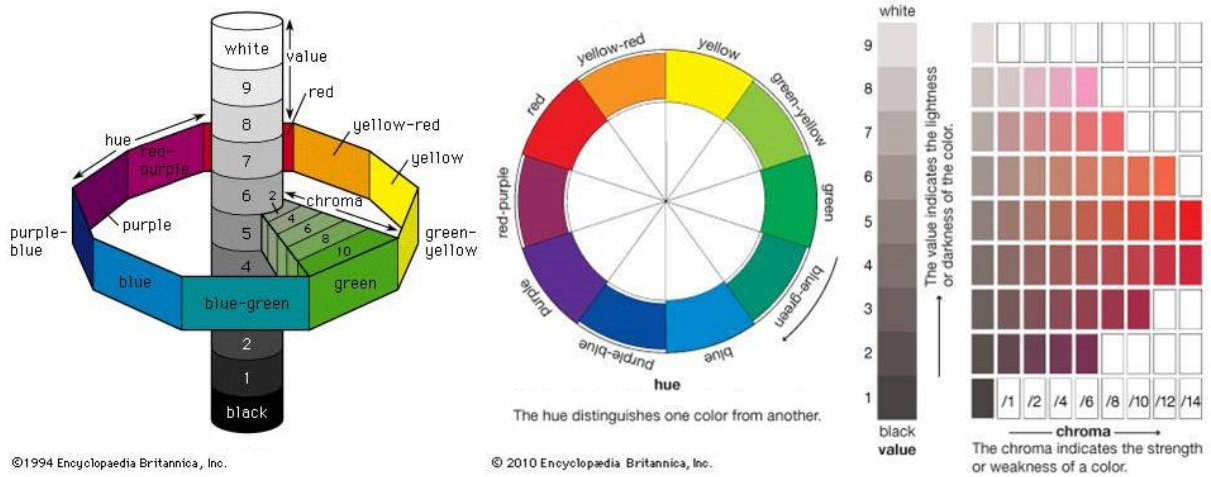


FIGURE A.2 – Illustrations du système de couleurs de Munsell.

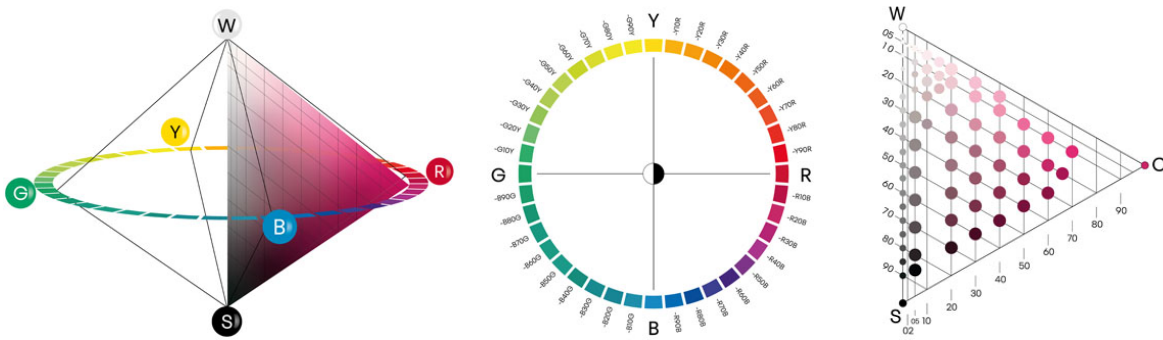


FIGURE A.3 – Illustrations du "système naturel de couleurs" (NCS).  
source : <https://nscolor.com/>

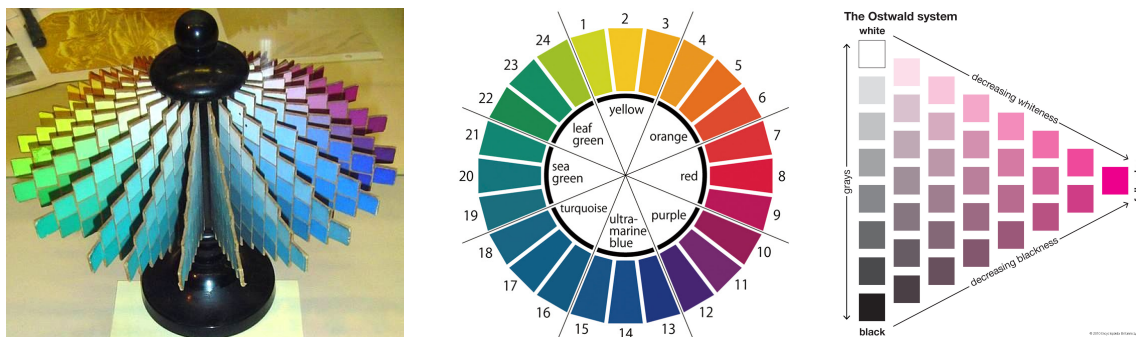


FIGURE A.4 – Illustrations du système de couleurs d'Ostwald.



FIGURE A.5 – Représentation 3D du système de couleurs de Munsell à partir de 10 ensembles de couleurs de même teinte.

## Annexe B Liste des publications

### Article de revue internationale

Robin DELÉARDE, Camille KURTZ et Laurent WENDLING. “Description and recognition of complex spatial configurations of object pairs with Force Banner 2D features”. In : *Pattern Recognition* 123 (2022), page 108410

### Articles dans les actes de conférences internationales

Robin DELÉARDE, Camille KURTZ, Philippe DEJEAN et Laurent WENDLING. “Force Banner for the recognition of spatial relations”. In : *International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pages 6065-6072

Robin DELÉARDE, Camille KURTZ, Philippe DEJEAN et Laurent WENDLING. “Segment My Object: A pipeline to extract segmented objects in images based on labels or bounding boxes”. In : *International Conference on Computer Vision Theory and Applications (VISAPP)*. 2021, pages 618-625

Mohamed-Hicham LEGHETTAS, Robin DELÉARDE, Camille KURTZ et Laurent WENDLING. “Combination of visual and semantic criteria for automated selection of region proposals in a bounding box”. In : *International Conference on Machine Vision (ICMV)*. Tome 12084. SPIE, 2021, pages 101-108

### Articles dans les actes de conférences nationales

Robin DELÉARDE, Camille KURTZ, Philippe DEJEAN et Laurent WENDLING. “Segmentation d’images à l’aide d’annotations faibles par combinaison de critères sémantique et géométriques”. In : *Extraction et Gestion des Connaissances (EGC) – Atelier "Apprentissage Profond: Théorie et Applications" (APTA)*. 2021, pages 2-14

Robin DELÉARDE, Camille KURTZ, Philippe DEJEAN et Laurent WENDLING. “Description et reconnaissance de relations spatiales avec le bandeau de forces”. In : *ORASIS*. 2021



# **Bibliographie**



## Références générales

- [1] Baptiste ABELOOS et Stéphane HERBIN. “Explaining object detectors: the case of transformer architectures”. In : *Workshop on Trustworthy Artificial Intelligence as a part of the ECML/PKDD*. 2022 <sup>5</sup>
- [2] John E. BALL, Derek T. ANDERSON et Chee Seng CHAN. “Comprehensive survey of deep learning in remote sensing: theories, tools, and challenges for the community”. In : *Journal of Applied Remote Sensing* 11.4 (sept. 2017), page 042609 <sup>4</sup>
- [3] Hila CHEFER, Shir GUR et Lior WOLF. “Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers”. In : *International Conference on Computer Vision (ICCV)*. IEEE, 2021, pages 397-406 <sup>5</sup>
- [4] Chaofan CHEN, Oscar LI, Daniel TAO, Alina BARNETT, Cynthia RUDIN et Jonathan K. SU. “This looks like that: deep learning for interpretable image recognition”. In : *Advances in Neural Information Processing Systems (NIPS)* 32 (2019) <sup>5</sup>
- [5] Robin DELÉARDE, Camille KURTZ, Philippe DEJEAN et Laurent WENDLING. “Description et reconnaissance de relations spatiales avec le bandeau de forces”. In : *ORASIS*. 2021 <sup>130</sup>
- [6] Robin DELÉARDE, Camille KURTZ, Philippe DEJEAN et Laurent WENDLING. “Force Banner for the recognition of spatial relations”. In : *International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pages 6065-6072 <sup>130</sup>
- [7] Robin DELÉARDE, Camille KURTZ, Philippe DEJEAN et Laurent WENDLING. “Segment My Object: A pipeline to extract segmented objects in images based on labels or bounding boxes”. In : *International Conference on Computer Vision Theory and Applications (VISAPP)*. 2021, pages 618-625 <sup>130</sup>
- [8] Robin DELÉARDE, Camille KURTZ, Philippe DEJEAN et Laurent WENDLING. “Segmentation d’images à l’aide d’annotations faibles par combinaison de critères sémantique et géométriques”. In : *Extraction et Gestion des Connaissances (EGC) – Atelier “Apprentissage Profond: Théorie et Applications” (APTA)*. 2021, pages 2-14 <sup>130</sup>
- [9] Robin DELÉARDE, Camille KURTZ et Laurent WENDLING. “Description and recognition of complex spatial configurations of object pairs with Force Banner 2D features”. In : *Pattern Recognition* 123 (2022), page 108410 <sup>130</sup>
- [10] Jia DENG, Wei DONG, Richard SOCHER, Li-Jia LI, Kai LI et Li FEI-FEI. “Imagenet: A large-scale hierarchical image database”. In : *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2009, pages 248-255 <sup>3</sup>
- [11] S. HERBIN et al. “Scene understanding from aerospace sensors: what can be expected?” In : *AerospaceLab* 4 (2012) <sup>4</sup>
- [12] Geoffrey E. HINTON, Alex KRIZHEVSKY et Sida D. WANG. “Transforming auto-encoders”. In : *International Conference on Artificial Neural Networks*. Springer, 2011, pages 44-51 <sup>5</sup>
- [13] Corentin KERVADEC, Theo JAUNET, Grigory ANTIPOV, Moez BACCOUCHE, Romain VUILLEMOT et Christian WOLF. “How transferable are reasoning patterns in VQA?” In : *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2021, pages 4207-4216 <sup>5</sup>
- [14] Alex KRIZHEVSKY, Ilya SUTSKEVER et Geoffrey E. HINTON. “Imagenet classification with deep convolutional neural networks”. In : *Advances in Neural Information Processing Systems (NIPS)*. 2012, pages 1097-1105 <sup>3</sup>
- [15] Yann LECUN, Léon BOTTOU, Yoshua BENGIO et Patrick HAFFNER. “Gradient-based learning applied to document recognition”. In : *Proceedings of the IEEE* 86.11 (1998), pages 2278-2324 <sup>3</sup>



- 
- [16] Yann LECUN et al. “Backpropagation applied to handwritten zip code recognition”. In : *Neural computation* 1.4 (1989), pages 541-551 <sup>3</sup>
  - [17] Mohamed-Hicham LEGHETTAS, Robin DELÉARDE, Camille KURTZ et Laurent WENDLING. “Combination of visual and semantic criteria for automated selection of region proposals in a bounding box”. In : *International Conference on Machine Vision (ICMV)*. Tome 12084. SPIE, 2021, pages 101-108 <sup>130</sup>
  - [18] Ahmed Samy NASSAR, Stefano D’ARONCO, Sébastien LEFÈVRE et Jan D. WEGNER. “GeoGraph: graph-based multi-view object detection with geometric cues end-to-end”. In : *European Conference on Computer Vision (ECCV)*. Springer, 2020, pages 488-504 <sup>133</sup>
  - [19] Ahmed Samy NASSAR, Sébastien LEFÈVRE et Jan Dirk WEGNER. “Simultaneous multi-view instance detection with learned geometric soft-constraints”. In : *International Conference on Computer Vision (ICCV)*. IEEE, 2019, pages 6559-6568 <sup>133</sup>
  - [20] Xiao Xiang ZHU et al. “Deep learning in remote sensing: a review”. In : *IEEE Geoscience and Remote Sensing Magazine* 5.4 (déc. 2017), pages 8-36 <sup>4</sup>

## Références de la Partie I

- [1] James F. ALLEN. “Maintaining knowledge about temporal intervals”. In : *Communications of the ACM* 26.11 (1983), pages 832-843 <sup>24, 36, 37</sup>
- [2] Jacob ANDREAS, Marcus ROHRBACH, Trevor DARRELL et Dan KLEIN. “Neural module networks”. In : *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016, pages 39-48 <sup>22, 23</sup>
- [3] Isabelle BLOCH. “Fuzzy relative position between objects in images: A morphological approach”. In : *International Conference on Image Processing (ICIP)*. Tome 96. 1996, pages 987-990 <sup>24, 42</sup>
- [4] Isabelle BLOCH. “Fuzzy relative position between objects in image processing: A morphological approach”. In : *IEEE Transactions on Pattern Analysis and Machine Intelligence* 21.7 (1999), pages 657-664 <sup>24, 42</sup>
- [5] Isabelle BLOCH. “Fuzzy spatial relationships for image processing and interpretation: A review”. In : *Image and Vision Computing* 23.2 (2005), pages 89-110 <sup>41</sup>
- [6] Isabelle BLOCH, Olivier COLLIOT et Roberto M. CESAR. “On the ternary spatial relation “between””. In : *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics* 36.2 (2006), pages 312-327 <sup>41</sup>
- [7] Isabelle BLOCH, Céline HUDELLOT et Jamal ATIF. “On the interest of spatial relations and fuzzy representations for ontology-based image interpretation”. In : *Advances In Pattern Recognition*. World Scientific, 2007, pages 15-25 <sup>12</sup>
- [8] Rajkumar BONDUGULA, Pascal MATSAKIS et James M. KELLER. “Force histograms and neural networks for human-based spatial relationship generalization”. In : *IASTED International Conference on Neural Networks and Computational Intelligence (NCI)*. ACTA Press, 2004, pages 185-190 <sup>42</sup>
- [9] Andrew R. BUCK, James M. KELLER et Marjorie SKUBIC. “A memetic algorithm for matching spatial configurations with the histograms of forces”. In : *IEEE Transactions on Evolutionary Computation* 17.4 (2013), pages 588-604 <sup>39</sup>
- [10] Yang CAO, Changhu WANG, Zhiwei LI, Liqing ZHANG et Lei ZHANG. “Spatial-bag-of-features”. In : *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. 2010, pages 3352-3359 <sup>70</sup>

- [11] Sung-Hyuk CHA. “Taxonomy of nominal type histogram distance measures”. In : *American Conference on Applied Mathematics*. 2008, pages 325-330 <sup>40</sup>
- [12] Sung-Hyuk CHA et Sargur N. SRIHARI. “On measuring the distance between histograms”. In : *Pattern Recognition* 35.6 (2002), pages 1355-1370 <sup>40</sup>
- [13] Shi-Kuo CHANG, Qing-Yun SHI et Cheng-Wen YAN. “Iconic indexing by 2-D strings”. In : *IEEE Transactions on Pattern Analysis and Machine Intelligence* 3 (1987), pages 413-428 <sup>25</sup>
- [14] Myung Jin CHOI, Joseph J. LIM, Antonio TORRALBA et Alan S. WILLSKY. “Exploiting hierarchical context on a large database of object categories”. In : *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2010, pages 129-136 <sup>23, 31</sup>
- [15] Jérémy CHOPIN. “Apprentissage profond et connaissances structurelles pour l’analyse d’images”. Thèse de doctorat. Université d’Angers (France), 2022 <sup>68</sup>
- [16] Jérémy CHOPIN, Jean-Baptiste FASQUEL, Harold MOUCHÈRE, Rozenn DAHYOT et Isabelle BLOCH. “QAP optimisation with reinforcement learning for faster graph matching in sequential semantic image analysis”. In : *International Conference on Pattern Recognition and Artificial Intelligence (ICPRAI)*. 2022 <sup>68</sup>
- [17] Michaël CLÉMENT. “Modélisation et apprentissage de relations spatiales pour la reconnaissance et l’interprétation d’images”. Thèse de doctorat. Université Sorbonne Paris Cité (Paris, France), 2017 <sup>38, 41, 67</sup>
- [18] Michaël CLÉMENT, Mickaël GARNIER, Camille KURTZ et Laurent WENDLING. “Color object recognition based on spatial relations between image layers”. In : *International Conference on Computer Vision Theory and Applications (VISAPP)*. Tome 1. 2015, pages 427-434 <sup>39, 68, 69, 71, 72, 76, 77</sup>
- [19] Michaël CLÉMENT, Camille KURTZ et Laurent WENDLING. “Bags of spatial relations and shapes features for structural object description”. In : *International Conference on Pattern Recognition (ICPR)*. IEEE, 2016, pages 1995-2000 <sup>69</sup>
- [20] Michaël CLÉMENT, Camille KURTZ et Laurent WENDLING. “Learning spatial relations and shapes for structural object description and scene recognition”. In : *Pattern Recognition* 84 (2018), pages 197-210 <sup>47, 51, 69</sup>
- [21] Michaël CLÉMENT, Camille KURTZ et Laurent WENDLING. “Fuzzy directional enlacement landscapes for the evaluation of complex spatial relations”. In : *Pattern Recognition* 101 (2020), page 107185 <sup>42</sup>
- [22] Michaël CLÉMENT, Adrien POULENARD, Camille KURTZ et Laurent WENDLING. “Directional enlacement histograms for the description of complex spatial configurations between Objects”. In : *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39.12 (2017), pages 2366-2380 <sup>41</sup>
- [23] Anthony G. COHN, Brandon BENNETT, John GOODAY et Nicholas Mark GOTTS. “Qualitative spatial representation and reasoning with the region connection calculus”. In : *GeoInformatica* 1.3 (1997), pages 275-316 <sup>16, 24, 36</sup>
- [24] Olivier COLLIOT, Oscar CAMARA et Isabelle BLOCH. “Integration of fuzzy spatial relations in deformable models – Application to brain MRI segmentation”. In : *Pattern Recognition* 39.8 (2006), pages 1401-1414 <sup>42</sup>
- [25] Donatello CONTE, Pasquale FOGGIA, Carlo SANSONE et Mario VENTO. “Thirty years of graph matching in pattern recognition”. In : *International journal of pattern recognition and artificial intelligence* 18.03 (2004), pages 265-298 <sup>68</sup>

- 
- [26] Bo DAI, Yuqi ZHANG et Dahua LIN. “Detecting visual relationships with deep relational networks”. In : *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pages 3298-3308 <sup>43</sup>
- [27] Sidi Mohammed Réda DEHAK. “Inférence quantitative des relations spatiales directionnelles”. Thèse de doctorat. Télécom ParisTech (Paris, France), 2002 <sup>41</sup>
- [28] Géraldine DEL MONDO. “Un modèle de graphe spatio-temporel pour représenter l’évolution d’entités géographiques”. Thèse de doctorat. Université de Bretagne Occidentale (Brest, France), 2011 <sup>70, 71</sup>
- [29] Géraldine DEL MONDO, John G. STELL, Christophe CLARAMUNT et Rémy THIBAUD. “A graph model for spatio-temporal evolution”. In : *Journal of Universal Computer Science* 16.11 (2010), pages 1452-1477 <sup>70, 71, 77</sup>
- [30] Robin DELÉARDE, Camille KURTZ, Philippe DEJEAN et Laurent WENDLING. “Description et reconnaissance de relations spatiales avec le bandeau de forces”. In : *ORASIS*. 2021 <sup>63</sup>
- [31] Robin DELÉARDE, Camille KURTZ, Philippe DEJEAN et Laurent WENDLING. “Force Banner for the recognition of spatial relations”. In : *International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pages 6065-6072 <sup>63</sup>
- [32] Robin DELÉARDE, Camille KURTZ et Laurent WENDLING. “Description and recognition of complex spatial configurations of object pairs with Force Banner 2D features”. In : *Pattern Recognition* 123 (2022), page 108410 <sup>63</sup>
- [33] Adrien DELIÈGE, Anthony CIOPPA et Marc VAN DROOGENBROECK. “An effective hit-or-miss layer favoring feature interpretation as learned prototypes deformations”. In : *AAAI Conference on Artificial Intelligence, Workshop on Network Interpretability for Deep Learning*. 2019 <sup>25</sup>
- [34] Agnes DESOLNEUX, Lionel MOISAN et Jean-Michel MOREL. *From gestalt theory to image analysis: a probabilistic approach*. Tome 34. Springer Science & Business Media, 2007 <sup>15</sup>
- [35] Xuewei DING, Yingwei PAN, Yehao LI, Ting YAO, Dan ZENG et Tao MEI. “Boosting relationship detection in images with multi-granular self-supervised learning”. In : *Transactions on Multimedia Computing, Communications, and Applications (TOMM)* (2022), pages – <sup>43</sup>
- [36] Max J. EGENHOFER et John R. HERRING. *Categorizing binary topological relations between regions, lines, and points in geographic databases*. Rapport technique. University of Maine, 1990 <sup>36, 39</sup>
- [37] Julien FAUQUEUR, Gabriel BROSTOW et Roberto CIPOLLA. “Assisted video object labeling by joint tracking of regions and keypoints”. In : *International Conference on Computer Vision (ICCV)*. IEEE, 2007, pages 1-7 <sup>29, 32</sup>
- [38] Bruno FERRARINI, Shoaib EHSAN, Adrien BARTOLI, Aleš LEONARDIS et Klaus D. McDONALD-MAIER. “Assessing capsule networks with biased data”. In : *Scandinavian Conference on Image Analysis*. Springer, 2019, pages 90-100 <sup>25</sup>
- [39] John FREEMAN. “The modelling of spatial relations”. In : *Computer Graphics and Image Processing* 4.2 (1975), pages 156-171 <sup>16, 24, 36</sup>
- [40] Christian FREKSA. “Qualitative spatial reasoning”. In : *Cognitive and linguistic aspects of geographic space*. Springer Netherlands, 1991, pages 361-372 <sup>16, 40</sup>
- [41] Mickaël GARNIER, Thomas HURTUT et Laurent WENDLING. “Object description based on spatial relations between level-sets”. In : *International Conference on Digital Image Computing Techniques and Applications (DICTA)*. IEEE, 2012, pages 1-7 <sup>68, 69, 71, 76</sup>
- [42] Ankit GOYAL, Kaiyu YANG, Dawei YANG et Jia DENG. “Rel3D: A minimally contrastive benchmark for grounding spatial relations in 3D”. In : *Advances in Neural Information Processing Systems (NIPS)* 33 (2020) <sup>23, 31, 43, 70</sup>

- [43] Hans Werner GUESGEN. *Spatial reasoning based on Allen's temporal logic*. International Computer Science Institute Berkeley, CA, 1989 <sup>36</sup>
- [44] D. S. GURU et P. NAGABHUSHAN. "Triangular spatial relationship: a new approach for spatial knowledge representation". In : *Pattern Recognition Letters* 22.9 (2001), pages 999-1006 <sup>66</sup>
- [45] M. HALDEKAR, A. GANESAN et T. OATES. "Identifying spatial relations in images using convolutional neural networks". In : *International Joint Conference on Neural Networks (IJCNN)*. 2017, pages 3593-3600 <sup>43</sup>
- [46] Daniel HERNÁNDEZ. "Relative representation of spatial knowledge: The 2-D case". In : *Cognitive and linguistic aspects of geographic space*. Springer, 1991, pages 373-385 <sup>36-39</sup>
- [47] Shawn HERSHEY et al. "CNN architectures for large-scale audio classification". In : *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2017, pages 131-135 <sup>55</sup>
- [48] Geoffrey E. HINTON, Alex KRIZHEVSKY et Sida D. WANG. "Transforming auto-encoders". In : *International Conference on Artificial Neural Networks*. Springer, 2011, pages 44-51 <sup>25</sup>
- [49] Nguyen Vu HOÀNG, Valérie GOUET-BRUNET, Marta RUKOZ et Maude MANOUVRIER. "Embedding spatial information into image content description for scene retrieval". In : *Pattern Recognition* 43.9 (2010), pages 3013-3024 <sup>66</sup>
- [50] Nguyen-Vu HOÀNG, Valérie GOUET-BRUNET et Marta RUKOZ. "Object detection and localization using a knowledge graph on spatial relationships". In : *International Conference on Multimedia and Expo (ICME)*. IEEE, 2013, pages 1-6 <sup>66, 70</sup>
- [51] M.W. IANDOLA F.N. and Moskewicz, K. ASHRAF, S. HAN, W.J. DALLY et K. KEUTZER. "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and 1MB model size". In : *CoRR* abs/1602.07360 (2016) <sup>55</sup>
- [52] Ray JACKENDOFF et Barbara LANDAU. "Spatial language and spatial cognition". In : *Bridges between psychology and linguistics*. Psychology Press, 2013, pages 157-182 <sup>16</sup>
- [53] M. JAZOULI, J. WADSWORTH et Pascal MATSAKIS. "Normalization of the histogram of forces". In : *International Conference on Pattern Recognition Applications and Methods (ICPRAM)*. 2019, pages 630-639 <sup>39, 40, 49, 72</sup>
- [54] Justin JOHNSON, Bharath HARIHARAN, Laurens VAN DER MAATEN, Li FEI-FEI, C. LAWRENCE ZITNICK et Ross GIRSHICK. "CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning". In : *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pages 2901-2910 <sup>22, 23</sup>
- [55] James M. KELLER et Xiaomei WANG. "Learning spatial relationships in computer vision". In : *IEEE International Conference on Fuzzy Systems (Fuzz-IEEE)*. Tome 1. IEEE, 1996, pages 118-124 <sup>42, 44</sup>
- [56] László T. KÓCZY. "On the description of relative position of fuzzy patterns". In : *Pattern Recognition Letters* 8.1 (1988), pages 21-28 <sup>24, 36, 41</sup>
- [57] Ranjay KRISHNA et al. "Visual Genome: Connecting language and vision using crowdsourced dense image annotation". In : *International Journal of Computer Vision* 123.1 (2017), pages 32-73 <sup>23, 43</sup>
- [58] Raghu KRISHNAPURAM, James M. KELLER et Yibing MA. "Quantitative analysis of properties and spatial relations of fuzzy image regions". In : *IEEE Transactions on Fuzzy Systems* 1.3 (1993), pages 222-233 <sup>41</sup>
- [59] Harold W KUHN. "The Hungarian method for the assignment problem". In : *Naval research logistics quarterly* 2.1-2 (1955), pages 83-97 <sup>67</sup>

- 
- [60] Alexander KUHNLE et Ann COPESTAKE. “ShapeWorld: A new test methodology for multimodal language understanding”. In : *CoRR* (2017) <sup>22, 23</sup>
- [61] Benjamin KUIPERS. “Modeling spatial knowledge”. In : *Cognitive science* 2.2 (1978), pages 129-153 <sup>16</sup>
- [62] Alina KUZNETSOVA et al. “The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale”. In : *CoRR* (2018) <sup>23</sup>
- [63] Tian LAN, Weilong YANG, Yang WANG et Greg MORI. “Image retrieval with structured object queries using latent ranking SVM”. In : *European Conference on Computer Vision (ECCV)*. Springer, 2012, pages 129-142 <sup>23, 31</sup>
- [64] Barbara LANDAU et Ray JACKENDOFF. ““What” and “where” in spatial language and spatial cognition”. In : *Behavioral and brain sciences* 16.2 (1993), pages 217-238 <sup>16</sup>
- [65] Aurélie LEBORGNE, Adrien MEYER, Henri GIRAUD, Florence LE BER et Stella MARC-ZWECKER. “Un graphe spatio-temporel pour modéliser l’évolution de parcelles agricoles”. In : *International Conference on Spatial Analysis and Geomatics (SAGEO)*. 2019 <sup>71</sup>
- [66] Gordon D. LOGAN et Daniel D. SADLER. “A computational analysis of the apprehension of spatial relations”. In : *Language, speech, and communication. Language and space*. MIT Press, 1996, pages 493-529 <sup>42</sup>
- [67] Nicolas LOMÉNIE et Daniel RACOCEANU. “Point set morphological filtering and semantic spatial configuration modeling: Application to microscopic image and bio-structure analysis”. In : *Pattern Recognition* 45.8 (2012), pages 2894-2911 <sup>41</sup>
- [68] Cewu LU, Ranjay KRISHNA, Michael BERNSTEIN et Li FEI-FEI. “Visual relationship detection with language priors”. In : *European Conference on Computer Vision (ECCV)*. Springer, 2016, pages 852-869 <sup>23, 31</sup>
- [69] Mateusz MALINOWSKI et Mario FRITZ. “A pooling approach to modelling spatial relations for image retrieval and annotation”. In : *CoRR* (2014) <sup>23, 42, 43</sup>
- [70] Jamal MALKI, Ei-Hadi ZAHZAH et Laurent MASCARILLA. “Indexation et recherche d’image fondées sur les relations spatiales entre objets”. In : *Traitement du signal* 18.4 (2002), pages 235-51 <sup>37, 39</sup>
- [71] Pascal MATSAKIS. “Relations spatiales structurelles et interprétation d’images”. Thèse de doctorat. Université Paul Sabatier (Toulouse, France), 1998 <sup>38</sup>
- [72] Pascal MATSAKIS. “Affine properties of the relative position PHI-descriptor”. In : *International Conference on Pattern Recognition (ICPR)*. IEEE, 2016, pages 1941-1946 <sup>39</sup>
- [73] Pascal MATSAKIS, James M. KELLER, Ozy SJAHPUTERA et Jonathon MARJAMAA. “The use of force histograms for affine-invariant relative position description”. In : *IEEE Transactions on Pattern Analysis and Machine Intelligence* 26.1 (2004), pages 1-18 <sup>40, 49, 73</sup>
- [74] Pascal MATSAKIS, James M. KELLER, Laurent WENDLING, Jonathan MARJAMAA et Ozy SJAHPUTERA. “Linguistic description of relative positions in images”. In : *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics* 31.4 (2001), pages 573-88 <sup>17, 24, 41, 42, 44, 46, 47, 51</sup>
- [75] Pascal MATSAKIS et Mohammad NAEEM. “Fuzzy models of topological relationships based on the phi-descriptor”. In : *IEEE International Conference on Fuzzy Systems (Fuzz-IEEE)*. 2016, pages 1096-1104 <sup>39</sup>
- [76] Pascal MATSAKIS, Mohammad NAEEM et Farhad RAHBARNIA. “Introducing the  $\Phi$ -descriptor – A most versatile relative position descriptor”. In : *International Conference on Pattern Recognition Applications and Methods (ICPRAM)*. 2015, pages 87-98 <sup>39</sup>



- [77] Pascal MATSAKIS, Jingbo NI et Xin WANG. “Object localization based on directional information: Case of 2D raster data”. In : *International Conference on Pattern Recognition (ICPR)*. IEEE, 2006, pages 142-146 <sup>42</sup>
- [78] Pascal MATSAKIS et Dennis NIKITENKO. “Combined extraction of directional and topological relationship information from 2D concave objects”. In : *Fuzzy Modeling with Spatial Information for Geographic Problems*. Springer Berlin Heidelberg, 2005, pages 15-40 <sup>39</sup>
- [79] Pascal MATSAKIS et Laurent WENDLING. “A new way to represent the relative position between areal objects”. In : *IEEE Transactions on Pattern Analysis and Machine Intelligence* 21.7 (1999), pages 634-643 <sup>24, 38, 44-47, 49, 50, 62, 72</sup>
- [80] Pascal MATSAKIS, Laurent WENDLING et JingBo NI. “A general approach to the fuzzy modeling of spatial relationships”. In : *Methods for Handling Imperfect Spatial Information*. 2010, pages 49-74 <sup>42, 44</sup>
- [81] Koji MIYAJIMA et Anca RALESCU. “Spatial organization in 2D segmented images: Representation and recognition of primitive spatial relations”. In : *Fuzzy Sets and Systems* 65.2 (1994), pages 225-236 <sup>38, 41</sup>
- [82] James MUNKRES. “Algorithms for the assignment and transportation problems”. In : *Journal of the society for industrial and applied mathematics* 5.1 (1957), pages 32-38 <sup>67</sup>
- [83] Mohammad NAEEM et Pascal MATSAKIS. “Relative position descriptors – A review”. In : *International Conference on Pattern Recognition Applications and Methods (ICPRAM)*. 2015, pages 286-295 <sup>38</sup>
- [84] Prem NAIR, Rohan DOSHI et Stefan KESELJ. *Pushing the limits of capsule networks*. 2018 <sup>25</sup>
- [85] Jingbo NI et Pascal MATSAKIS. “An equivalent definition of the histogram of forces: algorithmic implications”. In : *Pattern Recognition* 43.4 (2010), pages 1607-1617 <sup>40, 46, 49, 51</sup>
- [86] Jingbo NI, Pascal MATSAKIS et Lukasz WAWRZYNIAK. “Quantitative representation of the relative position between 3D objects”. In : *IASTED International Conference on Visualization, Imaging, and Image Processing (VIIP)*. 2004, pages 452-289 <sup>39</sup>
- [87] Ishan NIGAM, Chen HUANG et Deva RAMANAN. “Ensemble knowledge transfer for semantic segmentation”. In : *Winter Conference on Applications of Computer Vision (WACV)*. IEEE. 2018, pages 1499-1508 <sup>30, 33, 73</sup>
- [88] Danilo NUNES, Leonardo Anjoletto FERREIRA, Paulo E. SANTOS et Adam PEASE. “Representation and retrieval of images by means of spatial relations between objects”. In : *AAAI Spring Symposium: Combining Machine Learning with Knowledge Engineering*. 2019 <sup>43</sup>
- [89] Ali Ozgun OK, Caglar SENARAS et Baris YUKSEL. “Automated detection of arbitrarily shaped buildings in complex environments from monocular VHR optical satellite imagery”. In : *IEEE Transactions on Geoscience and Remote Sensing* 51.3 (2013), pages 1701-1717 <sup>42</sup>
- [90] Otávio AB PENATTI, Fernanda B. SILVA, Eduardo VALLE, Valerie GOUET-BRUNET et Ricardo da S. Torres TORRES. “Visual word spatial arrangement for image retrieval and classification”. In : *Pattern Recognition* 47.2 (2014), pages 705-720 <sup>70</sup>
- [91] Julia PEYRE, Ivan LAPTEV, Cordelia SCHMID et Josef SIVIC. “Weakly-supervised learning of visual relations”. In : *International Conference on Computer Vision (ICCV)*. IEEE, 2017, pages 5189-5198 <sup>43</sup>
- [92] P. PUNITHA et D. S. GURU. “Symbolic image indexing and retrieval by spatial similarity: An approach based on B-tree”. In : *Pattern Recognition* 41.6 (2008), pages 2068-2085 <sup>66</sup>

- 
- [93] R. QURESHI, J. RAMEL et Hubert CARDOT. “Graphic symbol recognition using flexible matching of attributed relational graphs”. In : *6th IASTED International Conference on VIII*. 2006, pages 383-388 <sup>68</sup>
- [94] Jathushan RAJASEGARAN, Vinoj JAYASUNDARA, Sandaru JAYASEKARA, Hirunima JAYASEKARA, Suranga SENEVIRATNE et Ranga RODRIGO. “Deepcaps: Going deeper with capsule networks”. In : *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2019, pages 10725-10733 <sup>25</sup>
- [95] David A. RANDELL, Zhan CUI et Anthony G. COHN. “A spatial logic based on regions and connection”. In : *International Conference on Principles of Knowledge Representation and Reasoning (KR)*. 1992, pages 165-176 <sup>16, 24, 36, 37</sup>
- [96] Azriel ROSENFELD et Avinash C KAK. *Digital picture processing*. Tome 2. Academic Press, 1982 <sup>16, 24, 72</sup>
- [97] Azriel ROSENFELD et Reinhard KLETTE. “Degree of adjacency or surroundedness”. In : *Pattern Recognition* 18.2 (1985), pages 169-177 <sup>41</sup>
- [98] Sara SABOUR, Nicholas FROSST et Geoffrey E HINTON. “Dynamic Routing Between Capsules”. In : *Advances in Neural Information Processing Systems (NIPS)*. Curran Associates, Inc., 2017, pages 3856-3866 <sup>25</sup>
- [99] Mohammad Amin SADEGHI et Ali FARHADI. “Recognition using visual phrases”. In : *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2011, pages 1745-1752 <sup>43</sup>
- [100] Nadeem SALAMAT et El-hadi ZAHZAH. “Spatio-temporal reasoning by combined topological and directional relations information”. In : *International Journal of Artificial Intelligence and Soft Computing* 3.2 (2012), pages 185-201 <sup>70</sup>
- [101] Jean-Pierre SALMON. “Reconnaissance de symboles complexes”. Thèse de doctorat. Institut National Polytechnique de Lorraine (Nancy, France), 2008 <sup>66</sup>
- [102] K. C. SANTOSH, Bart LAMIROY et Laurent WENDLING. “Symbol recognition using spatial relations”. In : *Pattern Recognition Letters* 33.3 (2012), pages 331-341 <sup>39, 69</sup>
- [103] K. C. SANTOSH, Bart LAMIROY et Laurent WENDLING. “Integrating vocabulary clustering with spatial relations for symbol recognition”. In : *International Journal on Document Analysis and Recognition (IJ DAR)* 17.1 (2014), pages 61-78 <sup>69</sup>
- [104] K. C. SANTOSH, Laurent WENDLING et Bart LAMIROY. “BoR: Bag-of-Relations for symbol retrieval”. In : *International Journal of Pattern Recognition and Artificial Intelligence (IJPRAI)* 28.6 (2014) <sup>69</sup>
- [105] Stefan SCHUBERT, Peer NEUBERT, Johannes PÖSCHMANN et Peter PRETZEL. “Circular convolutional neural networks for panoramic images and laser data”. In : *Intelligent Vehicles Symposium*. IEEE, 2019, pages 653-660 <sup>53</sup>
- [106] Ramprasaath R SELVARAJU, Michael COGSWELL, Abhishek DAS, Ramakrishna VEDANTAM, Devi PARIKH et Dhruv BATRA. “Grad-CAM: Visual explanations from deep networks via gradient-based localization”. In : *International Conference on Computer Vision (ICCV)*. IEEE, 2017, pages 618-626 <sup>61</sup>
- [107] Arnold W. M. SMEULDERS, Marcel WORRING, Simone SANTINI, Amarnath GUPTA et Ramesh JAIN. “Content-based image retrieval at the end of the early years”. In : *IEEE Transactions in Pattern Analysis and Machine Intelligence* 22.12 (2000), pages 1349-1380 <sup>12</sup>
- [108] Alane SUHR, Mike LEWIS, James YEH et Yoav ARTZI. “A corpus of natural language for visual reasoning”. In : *Annual Meeting of the Association for Computational Linguistics*. Tome 2. 2017, pages 217-223 <sup>22, 23</sup>



- [109] Salvatore TABBONE et Laurent WENDLING. “Color and grey level object retrieval using a 3D representation of force histogram”. In : *Image and Vision Computing* 21.6 (2003), pages 483-495 <sup>39</sup>
- [110] Salvatore TABBONE et Laurent WENDLING. “Retrieving images by content from strong relational graph matching”. In : *17th International Conference on Pattern Recognition (ICPR)*. Tome 2. IEEE, 2004, pages 951-954 <sup>25, 50, 68</sup>
- [111] Corina VĂDUVA, Inge GAVĂT et Mihai DATCU. “Latent Dirichlet allocation for spatial analysis of satellite images”. In : *IEEE Transactions on Geoscience and Remote sensing* 51.5 (2012), pages 2770-2786 <sup>13, 23, 32, 70</sup>
- [112] Maria Carolina VANEGAS, Isabelle BLOCH et Jordi INGLADA. “A fuzzy definition of the spatial relation “surround” - Application to complex shapes”. In : *European Society for Fuzzy Logic and Technology (EUSFLAT)*. 2011, pages 844-851 <sup>41</sup>
- [113] Laure VIEU. “Sémantique des relations spatiales et inférences spatio-temporelles: Une contribution à l’étude des structures formelles de l’espace en Langage Naturel”. Thèse de doctorat. Université Paul Sabatier (Toulouse, France), 1991 <sup>16</sup>
- [114] Xiaomei WANG et James M. KELLER. “Human-based spatial relationship generalization through neural/fuzzy approaches”. In : *Fuzzy Sets and Systems* 101.1 (1999), pages 5-20 <sup>42, 44, 51, 53</sup>
- [115] Yuhang WANG et Fillia MAKEDON. “R-Histogram: Quantitative representation of spatial relations for similarity-based image retrieval”. In : *ACM International Conference on Multimedia*. 2003, pages 323-326 <sup>39</sup>
- [116] Yuhang WANG, Fillia MAKEDON et Amit CHAKRABARTI. “R\*-Histograms: Efficient representation of spatial relations between objects of arbitrary topology”. In : *ACM International Conference on Multimedia*. 2004, pages 356-359 <sup>39</sup>
- [117] Laurent WENDLING. “Segmentation floue appliquée à la recherche d’objets dans les images numériques. Graphes relationnels et reconnaissance des formes. Application à la détection d’objets dans les images sur la base d’exemples.” Thèse de doctorat. Université Paul Sabatier (Toulouse, France), 1997 <sup>38, 67</sup>
- [118] Laurent WENDLING et Jacky DESACHY. “Isomorphism between strong fuzzy relational graphs based on k-formulae”. In : *Graph Based Representations in Pattern Recognition*. Springer, 1998, pages 63-71 <sup>68</sup>
- [119] Kaiyu YANG, Olga RUSSAKOVSKY et Jia DENG. “SpatialSense: An adversarially crowdsourced benchmark for spatial relation recognition”. In : *International Conference on Computer Vision (ICCV)*. IEEE, 2019, pages 2051-2060 <sup>20, 23, 28, 31, 43, 52, 54-56, 70</sup>
- [120] Bangpeng YAO et Li FEI-FEI. “Modeling mutual context of object and human pose in human-object interaction activities”. In : *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2010, pages 17-24 <sup>43</sup>

## Références du Chapitre 4

- [1] Jiwoon AHN, Sunghyun CHO et Suha KWAK. “Weakly supervised learning of instance segmentation with inter-pixel relations”. In : *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2019, pages 2209-2218 <sup>84</sup>
- [2] Pablo ARBELÁEZ, Jordi PONT-TUSET, Jonathan T. BARRON, Ferran MARQUES et Jitendra MALIK. “Multiscale combinatorial grouping”. In : *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2014, pages 328-335 <sup>84, 88</sup>

- 
- [3] Liang-Chieh CHEN, Yukun ZHU, George PAPANDEOU, Florian SCHROFF et Hartwig ADAM. “Encoder-decoder with atrous separable convolution for semantic image segmentation”. In : *European Conference on Computer Vision (ECCV)*. 2018, pages 801-818 <sup>89</sup>
- [4] Jifeng DAI, Kaiming HE et Jian SUN. “Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation”. In : *International Conference on Computer Vision (ICCV)*. 2015, pages 1635-1643 <sup>84, 86, 89</sup>
- [5] Robin DELÉARDE, Camille KURTZ, Philippe DEJEAN et Laurent WENDLING. “Segment My Object: A pipeline to extract segmented objects in images based on labels or bounding boxes”. In : *International Conference on Computer Vision Theory and Applications (VISAPP)*. 2021, pages 618-625 <sup>90</sup>
- [6] Robin DELÉARDE, Camille KURTZ, Philippe DEJEAN et Laurent WENDLING. “Segmentation d’images à l’aide d’annotations faibles par combinaison de critères sémantique et géométriques”. In : *Extraction et Gestion des Connaissances (EGC) – Atelier “Apprentissage Profond: Théorie et Applications” (APTA)*. 2021, pages 2-14 <sup>90</sup>
- [7] Ross GIRSHICK, Jeff DONAHUE, Trevor DARRELL et Jitendra MALIK. “Region-based convolutional networks for accurate object detection and segmentation”. In : *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38.1 (2015), pages 142-158 <sup>82</sup>
- [8] Matthieu GUILLAUMIN, Daniel KÜTTEL et Vittorio FERRARI. “Imagenet auto-annotation with segmentation propagation”. In : *International Journal of Computer Vision* 110.3 (2014), pages 328-348 <sup>84</sup>
- [9] Kaiming HE, Georgia GKIOXARI, Piotr DOLLÁR et Ross GIRSHICK. “Mask R-CNN”. In : *International Conference on Computer Vision (ICCV)*. 2017, pages 2961-2969 <sup>84, 87</sup>
- [10] Seunghoon HONG, Junhyuk OH, Honglak LEE et Bohyung HAN. “Learning transferrable knowledge for semantic segmentation with deep convolutional neural network”. In : *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016, pages 3204-3212 <sup>83</sup>
- [11] Cheng-Chun HSU, Kuang-Jui HSU, Chung-Chi TSAI, Yen-Yu LIN et Yung-Yu CHUANG. “Weakly supervised instance segmentation using the bounding box tightness prior”. In : *Advances in Neural Information Processing Systems (NIPS)*. 2019, pages 6586-6597 <sup>84, 85</sup>
- [12] Rosana EL JURDI, Caroline PETITJEAN, Paul HONEINE et Fahed ABDALLAH. “BB-UNet: U-Net with Bounding Box Prior”. In : *IEEE Journal of Selected Topics in Signal Processing* (2020) <sup>84</sup>
- [13] Anna KHOREVA, Rodrigo BENENSON, Jan HOSANG, Matthias HEIN et Bernt SCHIELE. “Simple does it: Weakly supervised instance and semantic segmentation”. In : *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pages 876-885 <sup>84, 85</sup>
- [14] Alexander KOLESNIKOV et Christoph H. LAMPERT. “Seed, expand and constrain: Three principles for weakly-supervised image segmentation”. In : *European Conference on Computer Vision (ECCV)*. 2016, pages 695-711 <sup>84</sup>
- [15] Mohamed-Hicham LEGHETTAS, Robin DELÉARDE, Camille KURTZ et Laurent WENDLING. “Combination of visual and semantic criteria for automated selection of region proposals in a bounding box”. In : *International Conference on Machine Vision (ICMV)*. Tome 12084. SPIE, 2021, pages 101-108 <sup>90</sup>
- [16] Victor LEMPITSKY, Pushmeet KOHLI, Carsten ROTHER et Toby SHARP. “Image segmentation with a bounding box prior”. In : *International Conference on Computer Vision (ICCV)*. 2009, pages 277-284 <sup>84, 85</sup>
- [17] Qizhu LI, Anurag ARNAB et Philip HS TORR. “Weakly-and semi-supervised panoptic segmentation”. In : *European Conference on Computer Vision (ECCV)*. 2018, pages 102-118 <sup>83</sup>

- [18] Tsung-Yi LIN et al. “Microsoft COCO: Common objects in context”. In : *European Conference on Computer Vision (ECCV)*. 2014, pages 740-755 <sup>82</sup>
- [19] Jonathan LONG, Evan SHELHAMER et Trevor DARRELL. “Fully convolutional networks for semantic segmentation”. In : *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2015, pages 3431-3440 <sup>82</sup>
- [20] George PAPANDEOU, Liang-Chieh CHEN, Kevin P. MURPHY et Alan L. YUILLE. “Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation”. In : *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2015, pages 1742-1750 <sup>83, 84</sup>
- [21] Lucas PASCAL, Xavier BOST et Benoit HUET. “Semantic and visual similarities for efficient knowledge transfer in CNN training”. In : *International Conference on Content-Based Multimedia Indexing (CBMI)*. 2019, pages 1-6 <sup>83, 85</sup>
- [22] Carsten ROTHER, Vladimir KOLMOGOROV et Andrew BLAKE. “GrabCut” interactive foreground extraction using iterated graph cuts”. In : *ACM Transactions on Graphics (TOG)* 23.3 (2004), pages 309-314 <sup>84, 88, 89</sup>
- [23] Ruoqi SUN, Xinge ZHU, Chongruo WU, Chen HUANG, Jianping SHI et Lizhuang MA. “Not all areas are equal: Transfer learning for semantic segmentation via hierarchical region selection”. In : *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pages 4360-4369 <sup>83</sup>
- [24] Yude WANG, Jie ZHANG, Meina KAN, Shiguang SHAN et Xilin CHEN. “Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation”. In : *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2020, pages 12275-12284 <sup>84, 89</sup>
- [25] Zifeng WU, Chunhua SHEN et Anton VAN DEN HENGEL. “Wider or deeper: Revisiting the ResNet model for visual recognition”. In : *Pattern Recognition* 90 (2019), pages 119-133 <sup>89</sup>
- [26] Kaiyu YANG, Olga RUSSAKOVSKY et Jia DENG. “SpatialSense: An adversarially crowdsourced benchmark for spatial relation recognition”. In : *International Conference on Computer Vision (ICCV)*. IEEE, 2019, pages 2051-2060 <sup>89</sup>
- [27] Yanzhao ZHOU, Yi ZHU, Qixiang YE, Qiang QIU et Jianbin JIAO. “Weakly supervised instance segmentation using class peak response”. In : *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2018, pages 3791-3800 <sup>84</sup>
- [28] Ganggao ZHU et Carlos A IGLESIAS. “Computing semantic similarity of concepts in knowledge graphs”. In : *IEEE Transactions on Knowledge and Data Engineering* 29.1 (2016), pages 72-85 <sup>87, 88</sup>

## Références du Chapitre 5

- [1] Jesus ANGULO et Jean SERRA. “Modelling and segmentation of colour images in polar representations”. In : *Image and Vision Computing* 25.4 (2007), pages 475-495 <sup>94, 97</sup>
- [2] Jesús ANGULO et Jean SERRA. “Traitements des images de couleur en représentation luminance/saturation/teinte par norme L1”. In : *Traitement du signal* 21.6 (2004), page 20 <sup>94, 97, 98, 104</sup>
- [3] Arindam BANERJEE, Inderjit S. DHILLON, Joydeep GHOSH, Suvrit SRA et Greg RIDGEWAY. “Clustering on the unit hypersphere using von Mises-Fisher distributions”. In : *Journal of Machine Learning Research* 6.9 (2005) <sup>99</sup>
- [4] Seema BANSAL et Deepak AGGARWAL. “Color image segmentation using CIE Lab color space using ant colony optimization”. In : *International Journal of Computer Applications* 29.9 (2011), pages 28-34 <sup>97</sup>

- 
- [5] Laurent BUSIN, Nicolas VANDENBROUCKE et Ludovic MACAIRE. "Color spaces and image segmentation". In : *Advances in imaging and electron physics* 151.1 (2008), page 1 <sup>95, 97</sup>
- [6] Thierry CARRON. "Segmentations d'images couleur dans la base Teinte-Luminance-Saturation: approche numérique et symbolique". Thèse de doctorat. Université de Chambéry (France), 1995 <sup>95</sup>
- [7] Thierry CARRON et Patrick LAMBERT. "Fuzzy color edge extraction by inference rules quantitative study and evaluation of performances". In : *International Conference on Image Processing (ICIP)*. Tome 2. IEEE. 1995, pages 181-184 <sup>95</sup>
- [8] Mehmet CELENK. "A color clustering technique for image segmentation". In : *Computer Vision, Graphics, and image processing* 52.2 (1990), pages 145-170 <sup>94, 96, 97</sup>
- [9] Sung-Hyuk CHA et Sargur N. SRIHARI. "On measuring the distance between histograms". In : *Pattern Recognition* 35.6 (2002), pages 1355-1370 <sup>99</sup>
- [10] Heng Da CHENG, X. H. JIANG, Ying SUN et Jingli WANG. "Color image segmentation: advances and prospects". In : *Pattern recognition* 34.12 (2001), pages 2259-2281 <sup>94, 97</sup>
- [11] Inderjit S. DHILLON et Dharmendra S. MODHA. "Concept decompositions for large sparse text data using clustering". In : *Machine learning* 42.1 (2001), pages 143-175 <sup>99</sup>
- [12] Farid GARCIA-LAMONT, Jair CERVANTES, Asdrúbal LÓPEZ et Lisbeth RODRIGUEZ. "Segmentation of images by color features: A survey". In : *Neurocomputing* 292 (2018), pages 1-27 <sup>95, 97</sup>
- [13] Richard S. HUNTER. "Photoelectric color difference meter". In : *Journal of the Optical Society of America (JOSA)* 48.12 (1958), pages 985-995 <sup>94</sup>
- [14] T.L. HUNTSHERGER, C.L. JACOBS et Robert L. CANNON. "Iterative fuzzy image segmentation". In : *Pattern recognition* 18.2 (1985), pages 131-138 <sup>95</sup>
- [15] Xu Ji, João F. HENRIQUES et Andrea VEDALDI. "Invariant information clustering for unsupervised image classification and segmentation". In : *International Conference on Computer Vision (ICCV)*. IEEE, 2019, pages 9865-9874 <sup>97</sup>
- [16] George H. JOBLOVE et Donald GREENBERG. "Color spaces for computer graphics". In : *Conference on Computer Graphics and Interactive Techniques*. 1978, pages 20-25 <sup>93</sup>
- [17] Zengwei JU, Jiazhong CHEN et Jingli ZHOU. "Image segmentation based on the HSI color space and an improved mean shift". In : (2013), pages 135-140 <sup>97</sup>
- [18] Asako KANEZAKI. "Unsupervised image segmentation by backpropagation". In : *International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2018, pages 1543-1547 <sup>97</sup>
- [19] John R. KENDER. *Saturation, hue, and normalized color: calculation, digitization effects, and use*. Rapport technique. Carnegie Mellon University in Pittsburgh, 1976 <sup>93</sup>
- [20] Wonjik KIM, Asako KANEZAKI et Masayuki TANAKA. "Unsupervised learning of image segmentation based on differentiable feature clustering". In : *Transactions on Image Processing* 29 (2020), pages 8055-8068 <sup>97</sup>
- [21] J.-H. LEE, B.-H. CHANG et S.-D. KIM. "Comparison of colour transformations for image segmentation". In : *Electronics Letters* 30.20 (1994), pages 1660-1661 <sup>94</sup>
- [22] Haim LEVKOWITZ et Gabor T. HERMAN. "GLHS: A generalized lightness, hue, and saturation color model". In : *CVGIP: Graphical Models and Image Processing* 55.4 (1993), pages 271-285 <sup>94</sup>
- [23] Xiaoyang LIU, Dean ZHAO, Weikuan JIA, Chengzhi RUAN, Shuping TANG et Tian SHEN. "A method of segmenting apples at night based on color and position information". In : *Computers and Electronics in Agriculture* 122 (2016), pages 118-123 <sup>97</sup>

- [24] Luca LUCCHESI et Sanjit K. MITRA. “Colour image segmentation: a state-of-the-art survey”. In : *Proceedings-Indian National Science Academy Part A* 67.2 (2001), pages 207-222 <sup>94</sup>
- [25] David L. MACADAM. *Color measurement: theme and variations*. Tome 27. Springer, 1985 <sup>94</sup>
- [26] J. MACQUEEN. “Classification and analysis of multivariate observations”. In : *Berkeley Symposium on Mathematical Statistics and Probability*. 1967, pages 281-297 <sup>96, 99, 105</sup>
- [27] Kanti V. MARDIA et Peter E. JUPP. *Directional statistics*. Tome 2. Wiley Online Library, 2000 <sup>98, 99</sup>
- [28] Ishan NIGAM, Chen HUANG et Deva RAMANAN. “Ensemble knowledge transfer for semantic segmentation”. In : *Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2018, pages 1499-1508 <sup>105</sup>
- [29] Ron OHLANDER, Keith PRICE et D. Raj REDDY. “Picture segmentation using a recursive region splitting method”. In : *Computer graphics and image processing* 8.3 (1978), pages 313-333 <sup>96</sup>
- [30] Ronald B. OHLANDER. *Analysis of natural scenes*. Rapport technique. Carnegie Mellon University in Pittsburgh, 1975 <sup>96</sup>
- [31] Yu-Ichi OHTA, Takeo KANADE et Toshiyuki SAKAI. “Color information for region segmentation”. In : *Computer graphics and image processing* 13.3 (1980), pages 222-241 <sup>94, 101</sup>
- [32] Nobuyuki OTSU. “A threshold selection method from gray-level histograms”. In : *IEEE transactions on systems, man, and cybernetics* 9.1 (1979), pages 62-66 <sup>96</sup>
- [33] Nikhil R. PAL et Sankar K. PAL. “A review on image segmentation techniques”. In : *Pattern recognition* 26.9 (1993), pages 1277-1294 <sup>94</sup>
- [34] Thrasyvoulos N. PAPPAS. “An adaptive clustering algorithm for image segmentation”. In : *IEEE Transactions on Signal Processing* 40.4 (1992), pages 901-914 <sup>96, 97</sup>
- [35] Thrasyvoulos N. PAPPAS et Nikil S. JAYANT. “An adaptive clustering algorithm for image segmentation”. In : *International Conference on Computer Vision (ICCV)*. IEEE, 1988, pages 310-311 <sup>96, 97</sup>
- [36] Sang Ho PARK, Il Dong YUN et Sang Uk LEE. “Color image segmentation based on 3-D clustering: morphological approach”. In : *Pattern Recognition* 31.8 (1998), pages 1061-1076 <sup>97</sup>
- [37] Eko PRASETYO, R. Dimas ADITYO, Nanik SUCIATI et Chastine FATICHAH. “Mango leaf image segmentation on HSV and YCbCr color spaces using Otsu thresholding”. In : *International Conference on Science and Technology-Computer (ICST)*. IEEE, 2017, pages 99-103 <sup>92, 96</sup>
- [38] T. PRIYA et P. KALAVATHI. “HSV based histogram thresholding technique for MRI brain tissue segmentation”. In : *International Symposium on Signal Processing and Intelligent Recognition Systems*. Springer, 2018, pages 322-333 <sup>92, 96</sup>
- [39] Sara SABOUR, Nicholas FROSST et Geoffrey E HINTON. “Dynamic Routing Between Capsules”. In : *Advances in Neural Information Processing Systems (NIPS)*. Curran Associates, Inc., 2017, pages 3856-3866 <sup>98, 99</sup>
- [40] Stefan SCHUBERT, Peer NEUBERT, Johannes PÖSCHMANN et Peter PRETZEL. “Circular convolutional neural networks for panoramic images and laser data”. In : *Intelligent Vehicles Symposium*. IEEE, 2019, pages 653-660 <sup>99</sup>
- [41] Jean SERRA. *Espaces couleur adaptés au traitement d'image*. Rapport technique. Centre de morphologie mathématique, École des Mines de Paris, 2003 <sup>94, 102</sup>
- [42] Jean SERRA. “Représentations de la couleur en coordonnées polaires adaptées au traitement d'images”. In : *École d'hiver sur l'image numérique couleur*. 2005 <sup>94, 103</sup>



- 
- [43] C. Y. SHEN, L. ZHOU, Z. W. TENG et J. X. WANG. "A new approach for recognition and position of traffic lights". In : *International Conference on Computer Information Systems and Industrial Applications*. Atlantis Press, 2015, pages 682-685 <sup>92, 96</sup>
  - [44] Alvy Ray SMITH. "Color gamut transform pairs". In : *ACM Siggraph Computer Graphics* 12.3 (1978), pages 12-19 <sup>93</sup>
  - [45] Alexander STREHL, Joydeep GHOSH et Raymond MOONEY. "Impact of similarity measures on web-page clustering". In : *AAAI workshop on artificial intelligence for web search*. Tome 58. 2000, page 64 <sup>99</sup>
  - [46] Johji TAJIMA. "Uniform color scale applications to computer graphics". In : *Computer Vision, Graphics, and Image Processing* 21.3 (1983), pages 305-325 <sup>102</sup>
  - [47] Kai TIAN, Jiu hao LI, Jiefeng ZENG, Asenso EVANS et Lina ZHANG. "Segmentation of tomato leaf images based on adaptive clustering number of K-means algorithm". In : *Computers and Electronics in Agriculture* 165 (2019), page 104962 <sup>97</sup>
  - [48] Shoji TOMINAGA. "Color image segmentation using three perceptual attributes". In : *Conference on Computer Vision and Pattern Recognition (CVPR)*. 1986, pages 628-630 <sup>96</sup>
  - [49] Shoji TOMINAGA. "Expansion of color images using three perceptual attributes". In : *Pattern recognition letters* 6.1 (1987), pages 77-85 <sup>100</sup>
  - [50] Shoji TOMINAGA. "Color classification of natural color images". In : *Color Research & Application* 17.4 (1992), pages 230-239 <sup>96, 101</sup>
  - [51] D.-C. TSENG et C.-H. CHANG. "Color segmentation using perceptual attributes". In : *International Conference on Pattern Recognition (ICPR)*. Tome III. Conference C: Image, Speech and Signal Analysis. IEEE, 1992, pages 228-231 <sup>97, 103, 104</sup>
  - [52] Sreenath Rao VANTARAM et Eli SABER. "Survey of contemporary trends in color image segmentation". In : *Journal of Electronic Imaging* 21.4 (2012), page 040901 <sup>94</sup>
  - [53] Martin VEJMEĽKA, Petr MUŠÍLEK, M. PALUŠ et Emil PELIKÁN. "K-means clustering for problems with periodic attributes". In : *International Journal of Pattern Recognition and Artificial Intelligence* 23.04 (2009), pages 721-743 <sup>98, 99, 101</sup>
  - [54] Arthur Robert WEEKS et G. Eric HAGUE. "Color segmentation in the HSI color space using the K-means algorithm". In : *Nonlinear Image Processing VIII*. Tome 3026. International Society for Optics et Photonics, 1997, pages 143-154 <sup>97</sup>
  - [55] Yixin YAN, Yongbin SHEN et Shengming LI. "Unsupervised color-texture image segmentation based on a new clustering method". In : *International Conference on New Trends in Information and Service Science*. IEEE, 2009, pages 784-787 <sup>97</sup>
  - [56] Wenzhu YANG, Sile WANG, Xiaolan ZHAO, Jingsi ZHANG et Jiaqi FENG. "Greenness identification based on HSV decision tree". In : *Information Processing in Agriculture* 2.3-4 (2015), pages 149-160 <sup>92, 96</sup>

## Références du Chapitre 6

- [1] Muna O. ALMASAWA, Lamiaa A. ELREFAEI et Kawthar MORIA. "A survey on deep learning-based person re-identification systems". In : *IEEE Access* 7 (2019), pages 175228-175247 <sup>118</sup>
- [2] SV ARUNA KUMAR, Ehsan YAGHOUBI, Abhijit DAS, B. S. HARISH et Hugo PROENÇA. "The P-DESTRE: A fully annotated dataset for pedestrian detection, tracking, and short/long-term re-identification from aerial devices". In : *IEEE Transactions on Information Forensics and Security* 16 (2020), pages 1696-1708 <sup>114, 116, 118, 119</sup>

- [3] Shi-Kuo CHANG, Qing-Yun SHI et Cheng-Wen YAN. “Iconic indexing by 2-D strings”. In : *IEEE Transactions on Pattern Analysis and Machine Intelligence* 3 (1987), pages 413-428 <sup>121</sup>
- [4] Chaofan CHEN, Oscar LI, Daniel TAO, Alina BARNETT, Cynthia RUDIN et Jonathan K. SU. “This looks like that: deep learning for interpretable image recognition”. In : *Advances in Neural Information Processing Systems (NIPS)* 32 (2019) <sup>117</sup>
- [5] Michaël CLÉMENT, Mickaël GARNIER, Camille KURTZ et Laurent WENDLING. “Color object recognition based on spatial relations between image layers”. In : *International Conference on Computer Vision Theory and Applications (VISAPP)*. Tome 1. 2015, pages 427-434 <sup>121</sup>
- [6] Jian DONG, Qiang CHEN, Xiaohui SHEN, Jianchao YANG et Shuicheng YAN. “Towards unified human parsing and pose estimation”. In : *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2014, pages 843-850 <sup>120</sup>
- [7] Jianlong FU, Heliang ZHENG et Tao MEI. “Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition”. In : *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pages 4438-4446 <sup>117</sup>
- [8] Mengran GOU, Ziyang WU, Angels RATES-BORRAS, Octavia CAMPS et Richard J. RADKE. “A systematic evaluation and benchmark for person re-identification: Features, metrics, and datasets”. In : *Transactions on Pattern Analysis and Machine Intelligence* 41.3 (2018), pages 523-536 <sup>118</sup>
- [9] Aleksei GRIGOREV, Zhihong TIAN, Seungmin RHO, Jianxin XIONG, Shaohui LIU et Feng JIANG. “Deep person re-identification in UAV images”. In : *EURASIP Journal on Advances in Signal Processing* 1 (2019), pages 1-10 <sup>116, 118</sup>
- [10] Chunhui GU, Joseph J. LIM, Pablo ARBELÁEZ et Jitendra MALIK. “Recognition using regions”. In : *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2009, pages 1030-1037 <sup>117</sup>
- [11] Stéphane HERBIN. “Fine-grained recognition by sequential hypothesis rejection and foveated vision on parts”. In : *ECCV workshop on Parts and Attributes*. 2014, pages 68-69 <sup>117, 118</sup>
- [12] Geoffrey E. HINTON, Alex KRIZHEVSKY et Sida D. WANG. “Transforming auto-encoders”. In : *International Conference on Artificial Neural Networks*. Springer, 2011, pages 44-51 <sup>118</sup>
- [13] Jan HOSANG, Rodrigo BENENSON, Piotr DOLLÁR et Bernt SCHIELE. “What makes for effective detection proposals?” In : *Transactions on Pattern Analysis and Machine Intelligence* 38.4 (2015), pages 814-830 <sup>117</sup>
- [14] Harold W KUHN. “The Hungarian method for the assignment problem”. In : *Naval research logistics quarterly* 2.1-2 (1955), pages 83-97 <sup>121</sup>
- [15] Ryan LAYNE, Timothy M. HOSPEDALES et Shaogang GONG. “Investigating open-world person re-identification using a drone”. In : *European Conference on Computer Vision (ECCV)*. Springer, 2014, pages 225-240 <sup>116, 118</sup>
- [16] Kevin LIN, Lijuan WANG, Kun LUO, Yinpeng CHEN, Zicheng LIU et Ming-Ting SUN. “Cross-domain complementary learning using pose for multi-person part segmentation”. In : *Transactions on Circuits and Systems for Video Technology* 31.3 (2020), pages 1066-1078 <sup>116, 120</sup>
- [17] Tsung-Yu LIN, Aruni ROYCHOWDHURY et Subhransu MAJI. “Bilinear CNN models for fine-grained visual recognition”. In : *International Conference on Computer Vision (ICCV)*. IEEE, 2015, pages 1449-1457 <sup>117</sup>
- [18] Xiangtan LIN, Pengzhen REN, Yun XIAO, Xiaojun CHANG et Alex HAUPTMANN. “Person Search Challenges and Solutions: A Survey”. In : *International Joint Conference on Artificial Intelligence (IJCAI)*. 2021, pages 1-10 <sup>118</sup>
- [19] James MUNKRES. “Algorithms for the assignment and transportation problems”. In : *Journal of the society for industrial and applied mathematics* 5.1 (1957), pages 32-38 <sup>121</sup>



- 
- [20] Ishan NIGAM, Chen HUANG et Deva RAMANAN. “Ensemble knowledge transfer for semantic segmentation”. In : *Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2018, pages 1499-1508 <sup>116, 122</sup>
- [21] George PAPANDEOU, Tyler ZHU, Liang-Chieh CHEN, Spyros GIDARIS, Jonathan TOMPSON et Kevin MURPHY. “Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model”. In : *European Conference on Computer Vision (ECCV)*. 2018, pages 269-286 <sup>120</sup>
- [22] Olivier RISSER-MAROIX. “Similarité visuelle et apprentissage de représentations”. Thèse de doctorat. Université Paris Cité (France), 2022 <sup>114</sup>
- [23] Sara SABOUR, Nicholas FROSST et Geoffrey E HINTON. “Dynamic Routing Between Capsules”. In : *Advances in Neural Information Processing Systems (NIPS)*. Curran Associates, Inc., 2017, pages 3856-3866 <sup>118</sup>
- [24] Lars SOMMER, Andreas SPECKER et Arne SCHUMANN. “Deep learning based person search in aerial imagery”. In : *Automatic Target Recognition XXXI*. Tome 11729. International Society for Optics et Photonics, 2021 <sup>118</sup>
- [25] Malik SOUDED. “People detection, tracking and re-identification through a video camera network”. Thèse de doctorat. Université Nice Sophia Antipolis, 2013 <sup>118</sup>
- [26] Andreas SPECKER, Lennart MORITZ et Lars SOMMER. “Deep learning-based video analysis pipeline for person detection and re-identification in aerial imagery”. In : *Counterterrorism, Crime Fighting, Forensics, and Surveillance Technologies V*. Tome 11869. SPIE, 2021, pages 86-95 <sup>118</sup>
- [27] Tinne TUYTELAARS et Krystian MIKOLAJCZYK. “Local invariant feature detectors: a survey”. In : *Foundations and trends® in computer graphics and vision 3.3* (2008), pages 177-280 <sup>117</sup>
- [28] Fangting XIA, Peng WANG, Xianjie CHEN et Alan L. YUILLE. “Joint multi-person pose estimation and semantic part segmentation”. In : *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pages 6769-6778 <sup>120</sup>
- [29] Tianjun XIAO, Yichong XU, Kuiyuan YANG, Jiaying ZHANG, Yuxin PENG et Zheng ZHANG. “The application of two-level attention models in deep convolutional neural network for fine-grained image classification”. In : *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2015, pages 842-850 <sup>117</sup>
- [30] Bangpeng YAO, Aditya KHOSLA et Li FEI-FEI. “Combining randomization and discrimination for fine-grained image categorization”. In : *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2011, pages 1577-1584 <sup>117</sup>
- [31] Dengsheng ZHANG et Guojun LU. “Review of shape representation and description techniques”. In : *Pattern recognition 37.1* (2004), pages 1-19 <sup>117</sup>
- [32] Ning ZHANG, Jeff DONAHUE, Ross GIRSHICK et Trevor DARRELL. “Part-based R-CNNs for fine-grained category detection”. In : *European Conference on Computer Vision (ECCV)*. Springer, 2014, pages 834-849 <sup>117</sup>
- [33] Shizhou ZHANG et al. “Person re-identification in aerial imagery”. In : *IEEE Transactions on Multimedia 23* (2020), pages 281-291 <sup>116, 118</sup>
- [34] Heliang ZHENG, Jianlong FU, Tao MEI et Jiebo LUO. “Learning multi-attention convolutional neural network for fine-grained image recognition”. In : *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pages 5209-5217 <sup>117</sup>

## Références des Annexes

- [1] Isaac NEWTON. *Opticks*. 1704 <sup>137</sup>
- [2] Thomas YOUNG. “II. The Bakerian lecture. On the theory of light and colours”. In : *Philosophical Transactions of the Royal Society of London* 92 (1802), pages 12-48 <sup>137</sup>
- [3] Johann Wolfgang GOETHE. *Zur Farbenlehre*. 1810 <sup>138</sup>
- [4] P. O. RUNGE. “Die Farben-Kugel, oder Construction des Verhältnisses aller Farben zueinander [Color Sphere, the construction of all mutual relations of colors]”. In : *Perthes* (1810) <sup>137</sup>
- [5] Michel E. CHEVREUL. *De la loi de contraste simultané des couleurs*. Tome 1. Pitois-Levrault, 1839 <sup>138</sup>
- [6] Hermann GRASSMANN. “Zur theorie der farbenmischung”. In : *Annalen der Physik* 165.5 (1853), pages 69-84 <sup>138</sup>
- [7] James Clerk MAXWELL. “XVIII — Experiments on colour, as perceived by the eye, with remarks on colour-blindness”. In : *Earth and Environmental Science Transactions of the Royal Society of Edinburgh* 21.2 (1857), pages 275-298 <sup>138</sup>
- [8] Ewald HERING. *Zur Lehre vom Lichtsinne: sechs Mittheilungen an die Kaiserl. Akademie der Wissenschaften in Wien*. C. Gerold’s Sohn, 1878 <sup>138</sup>
- [9] Ogden Nicholas ROOD. *Modern chromatics: With applications to art and industry*. Tome 27. C. Kegan Paul, 1879 <sup>138</sup>
- [10] H. VON HELMHOLTZ. “Handbuch der physiologischen optik”. In : *Leipzig Dritter Abschnitt, Voss* (1886), pages 204-205 <sup>137</sup>
- [11] Wilhelm OSTWALD. *Die farbenfibel*. Verlag Unesma, 1917 <sup>137, 138</sup>
- [12] Albert Henry MUNSELL. *A color notation*. Munsell color company, 1919 <sup>137, 138</sup>
- [13] Soumya CHATTERJEE et Edward M. CALLAWAY. “Parallel colour-opponent pathways to primary visual cortex”. In : *Nature* 426.6967 (2003), pages 668-671 <sup>138</sup>
- [14] Rolf G. KUEHNI. “Forgotten pioneers of color order. Part I: Gaspard Grégoire (1751–1846)”. In : *Color Research & Application, endorsed by the Inter-Society Color Council* 33.1 (2008), pages 5-9 <sup>137, 138</sup>
- [15] Andrew STOCKMAN et David H. BRAINARD. “Color vision mechanisms”. In : *The Optical Society of America handbook of optics* 3 (2010), pages 11-1 <sup>138</sup>