Description et reconnaissance de relations spatiales avec le bandeau de forces

Robin Deléarde, Camille Kurtz, Philippe Dejean, Laurent Wendling

Description et reconnaissance de relations spatiales avec le bandeau de forces
Description and recognition of spatial relations with the force banner

R. Deléarde[1,2] C. Kurtz[1] P. Dejean[2] L. Wendling[1]

[1] LIPADE, Université de Paris – Paris, France
[2] Magellium, groupe Artal – Toulouse, France

robin.delearde@etu.u-paris.fr

Résumé

Un défi dans la compréhension de scènes est la prise en compte des relations spatiales entre les objets. Plusieurs descripteurs existent déjà pour cela, 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 complexités de configuration variables. Les résultats expérimentaux obtenus mettent en évidence l'intérêt de cette approche, en particulier pour des configurations spatiales complexes.*

Mots-clefs

Analyse d'images, Compréhension de scènes, Relations spatiales, Histogramme de forces

Abstract

*A challenge in scene understanding is the handling of spatial relations between objects. Several descriptors already exist for that, such as the force histogram which is a typical example of relative position descriptor. By computing the interaction between objects for a given force in all the directions, it gives a good overview of the configuration, and it has useful properties that can make it invariant to the 2D viewpoint. Considering that using complementary forces should improve the description of complex spatial configurations, we propose to extend the force histogram to a panel of forces so as to make it a more complete descriptor. This gives a 2D descriptor that we called "*force banner*", which can be used as input of a Convolutional Neural Network (CNN), benefiting from their powerful performances, or reduced into more compact spatial features to use them in another system. As an illustration of its ability to describe spatial configurations, we used it to solve a classification problem aiming to discriminate simple spatial relations, but with variable configuration complexities. Experimental results obtained highlight the interest of this approach, in particular for complex spatial configurations.*

Keywords

Image analysis, Scene understanding, Spatial relations, Force Histogram

1 Introduction

Depuis plusieurs années, il est acquis qu'il est nécessaire d'exploiter les relations spatiales entre les objets pour mieux comprendre les scènes. Ces informations spatiales devraient compléter les caractéristiques traditionnelles basées sur le contour, la géométrie, la texture, la couleur, 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. Leur modélisation peut être très utile pour de nombreuses tâches de vision par ordinateur, des tâches de base comme la comparaison et la reconnaissance d'objets ou de scènes, à des tâches appliquées comme la recherche d'images basée sur le contenu, le suivi d'objets, la navigation, etc. 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.

Cela conduit au problème fondamental de la prise en compte des relations spatiales entre paires d'objets : étant donné deux objets représentés dans une image, comment extraire et décrire efficacement leur configuration spatiale ?

Dans cet article, nous proposons une représentation qui décrit la forme des objets et leur distance 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 [23], 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 par d'autres modèles de classification pour améliorer leur compréhension spatiale de la scène, soit tel quel avec un CNN, soit avec un autre type de modèle en le rendant plus compact si besoin.

2 État de l'art

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 premières recherches sur la description des relations spatiales entre objets sont celles de Freeman, qui dans les années 1970 a proposé une catégorisation en 13 relations spatiales qualitatives en langage naturel [14]. D'autres catégorisations similaires ont été proposées ensuite, comme le *Region Connection Calculus* (RCC) [27, 11] qui est toujours la référence dans ce domaine, ou plusieurs solutions utilisant des projections des objets sur une seule dimension, comme la solution de [17] basée sur les intervalles temporels d'Allen [1]. Ces relations spatiales peuvent être directionnelles ("*à gauche de*", "*au-dessus*", etc.) ou topologiques ("*à l'intérieur*", "*à l'extérieur*", "*tangent*", etc.). 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 » [14]. 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. C'est pourquoi de nombreux auteurs ont assimilé les objets $2D$ à des entités très élémentaires telles qu'un point (centroïde) ou un rectangle (englobant). La procédure est

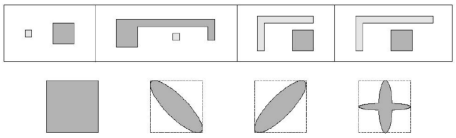

FIGURE 1 – Exemples de configurations spatiales entre deux objets binaires. Dans certaines situations, les approches centroïde ou rectangle englobant peuvent entraîner des ambiguïtés.

commode, mais ne peut pas fournir une modélisation satisfaisante, comme l'a souligné Rosenfeld en 1985 [28] (voir Figure 1). Ce n'est qu'après plusieurs années que les relations floues ont pu réellement être exploitées, par exemple avec les paysages flous [2]. Ce modèle basé sur des opérations morphologiques a permis d'obtenir des résultats intéressants dans diverses applications telles que la reconnaissance faciale [6], la segmentation cérébrale par IRM [12] ou encore la reconnaissance de texte [13].

Une approche parallèle consiste à utiliser les mesures quantitatives elles-mêmes comme descripteurs de la configuration spatiale. On parle alors de descripteurs de position relative lorsque le but est de décrire complètement la configuration, en utilisant une combinaison de plusieurs mesures. De cette manière, la position relative d'un objet par rapport à un autre peut avoir une représentation qui lui est propre, à partir de laquelle il est possible de déduire des évaluations de différentes relations spatiales [32, 19].

Un exemple célèbre de descripteur de position relative est l'histogramme de forces, proposé dans [23], qui est une généralisation de l'histogramme d'angles [24]. En calculant l'interaction entre les objets pour chaque direction, il combine l'aspect directionnel et des mesures de distance, tout en tenant compte de la forme et de la dimension des objets. Sa définition mathématique est donnée dans la Section 3, où certaines propriétés intéressantes sont également rapportées. Les histogrammes de force sont impliqués dans plusieurs domaines d'application, tels que la mise en correspondance de scènes [5] ou la recherche d'images basée sur le contenu [30, 7].

Récemment, une extension de l'histogramme de forces utilisant plusieurs histogrammes différents a été proposée [22, 21], afin de donner une description encore plus précise de la configuration spatiale. Cette représentation appelée ϕ-descripteur est basée sur une catégorisation selon les intervalles temporels d'Allen [1], et fournit un cadre générique pour évaluer n'importe quelle relation spatiale usuelle, en extrayant un ensemble d'opérateurs dédiés. Enfin, d'autres descripteurs ont également été proposés pour modéliser des relations spatiales plus spécifiques, par exemple pour les relations "*à travers*", "*le long de*", "*entre*" [4, 18], "*entouré par*" [31], ou encore "*enlacé par*" [10], avec une stratégie inspirée de l'histogramme de forces pour cette dernière.

Ainsi, nous pouvons distinguer deux types de descripteurs de configuration spatiale, avec une dualité importante [25, 3] : (1) les descripteurs de relations spatiales, décrivant la configuration selon une ou plusieurs relations spatiales qualitatives, où chaque relation peut 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, et qui peuvent être utilisés pour évaluer différentes relations spatiales.

Généralement calculés entre deux objets ou parties d'objets, tous ces descripteurs peuvent alors être associés à une décomposition de scène ou d'objet pour 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. Par exemple, [29] a introduit des représentations par « sacs de relations », combinant les primitives visuelles issues de plusieurs paires de relations spatiales. Et [8, 9] ont introduit la décomposition en histogrammes de force (FHD), un descripteur hiérarchique basé sur des graphes qui permet de caractériser les relations spatiales et les informations de forme des sous-parties structurelles des objets. Dans cette approche, un nouveau cadre de sacs de relations est utilisé pour produire des caractéristiques structurelles discriminantes adaptées à des tâches de classification d'objets, avec l'avantage d'être compatible avec les approches traditionnelles de « sacs de caractéristiques », permettant des représentations hybrides qui rassemblent des caractéristiques structurelles et locales.

3 Méthodologie

L'histogramme de forces a apporté une avancée dans le domaine de la compréhension de scènes, en permettant de décrire plus efficacement des configurations spatiales variées, et a pu être utilisé pour de nombreuses applications. Il dépend d'un paramètre de force, qui modélise différentes perceptions de la relation entre les objets, des forces répulsives aux forces attractives [23]. Dans ce contexte, il semble utile de combiner plusieurs histogrammes afin d'obtenir une meilleure compréhension de la relation, comme initié dans [19] avec deux forces spécifiques : F_0 and 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) [23].

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 [32] qui utilise l'histogramme d'angles.

Nous développons deux pistes principales dans cet article :

- 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 donnant certaines propriétés théoriques qu'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.

3.1 Notions sur l'histogramme de forces

Les histogrammes de forces visent à évaluer et à caractériser la configuration spatiale directionnelle entre objets binaires dans les images. Ce modèle a été introduit dans [23], reposant sur la définition d'une force d'interaction entre les objets. Concrètement, l'histogramme de forces est calculé en intégrant une force élémentaire fonction de la distance entre points. Étant donné deux points situés à une distance d l'un de l'autre, leur force d'attraction est $\varphi_r(d) = \frac{1}{d^r}$, où r caractérise le type de force utilisée : attractive ($r > 0$), répulsive ($r < 0$) ou constante ($r = 0$).

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 θ, D^{θ}_{IJ} leur distance et $|.|$ la longueur d'un segment. La force d'attraction f_r du segment I par rapport au segment J est donnée par :

$$f_r(I,J) = \int_{D^{\theta}_{IJ}+|J|}^{|I|+D^{\theta}_{IJ}+|J|} \int_0^{|J|} \varphi_r(u-v)\,\mathrm{d}v\,\mathrm{d}u. \quad (1)$$

Étant donné deux objets binaires A et B, une droite orientée de direction θ 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 $F_r(\theta, \mathcal{C}_A, \mathcal{C}_B) = \sum_{I\in\mathcal{C}_A}\sum_{J\in\mathcal{C}_B} f_r(I,J)$. Alors, l'ensemble des droites parallèles de direction θ 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 θ. La Figure 2 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

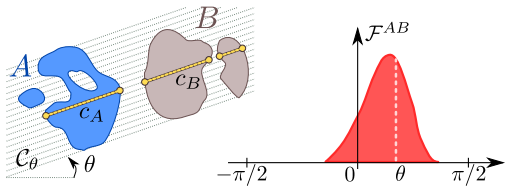

FIGURE 2 – Illustration du calcul de l'histogramme de forces. La force d'attraction entre A et B selon la direction θ est l'intégrale des forces calculées sur les coupes longitudinales(C_A, C_B) [23].

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 A et lui-même (*i.e.*, si nous calculons $\mathcal{F}^{AA}$), alors il peut être considéré comme un descripteur de forme de l'objet A.

3.2 Vers le bandeau de forces

Si l'on considère l'histogramme de forces, deux niveaux de forces sont largement utilisés dans la littérature pour évaluer la relation spatiale entre deux objets [23, 19] :

- $r = 0$ 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 indépendante à des changements d'échelle appliqués aux deux objets, après normalisation.

L'évaluation de 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. Un tel comportement peut être corrigé en considérant F_2^{AB}, qui se concentre sur des vues rapprochées entre 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 [19].
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 [9] (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 3.4) pour déduire les relations spatiales entre objets, ou comme primitives spatiales dans un système de reconnaissance, après une étape de compression éventuellement.

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

Soient A et B deux objets binaires et soient r et θ 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 :

$$\begin{aligned} \mathcal{FB}^{AB} : [0, 2\pi[\times [r_s, r_e] &\rightarrow \mathbb{R}_+ \\ (\theta, r) &\mapsto \mathcal{F}_r^{AB}(\theta) \end{aligned} \tag{2}$$

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 $\Theta = \{\theta_1, \theta_2, \ldots, \theta_{|\Theta|}\}$ 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 $R = \{r_s, r_s + \delta_r, \ldots, r_e\}$ un ensemble de niveaux de forces entre $r_s \in \mathbb{R}$ et $r_e \in \mathbb{R}$, avec un pas constant δ_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 :

$$d\mathcal{FB}^{AB} = \begin{vmatrix} \mu_{\theta_0, r_s} & \cdots & \mu_{\theta_{0+i\delta_\theta}, r_s} & \cdots & \mu_{\theta_{2\pi-\delta_\theta}, r_s} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \mu_{\theta_0, r_j} & \cdots & \mu_{\theta_{0+i\delta_\theta}, r_j} & \cdots & \mu_{\theta_{2\pi-\delta_\theta}, r_j} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \mu_{\theta_0, r_e} & \cdots & \mu_{\theta_{0+i\delta_\theta}, r_e} & \cdots & \mu_{\theta_{2\pi-\delta_\theta}, r_e} \end{vmatrix} \tag{3}$$

et

$$\mu_{\theta_i, r_j} = \frac{\mathcal{F}_{r_j}^{AB}(\theta_i)}{||\mathcal{F}_{r_j}^{AB}(.)||} \tag{4}$$

avec $r_j = r_{s+j\delta_r}$.
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 θ. 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 3 donne quelques exemples de bandeaux de forces discrets provenant des différents jeux de données que nous avons utilisés dans nos expérimentations.
Par héritage de l'histogramme de forces, le bandeau de forces a plusieurs propriétés bien utiles. L'une des propriétés les plus intéressantes est qu'il peut facilement prendre en compte 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 correspondrait à une variation du

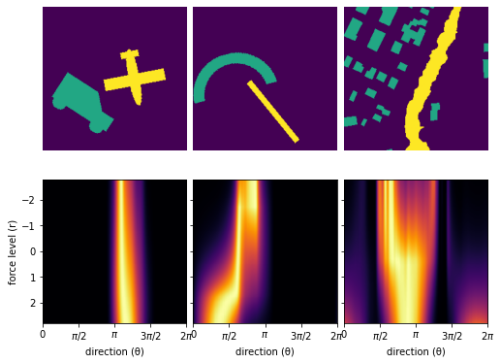

FIGURE 3 – (1^e ligne) Exemples issus du jeu de données de formes binaires (de gauche à droite) : objets synthétiques, formes géométriques synthétiques, objets (rivière et maisons) d'une image SIG segmentée, objets (route et bus) d'une image naturelle segmentée. Le référent est en jaune tandis que l'argument est en vert. (2^e ligne) Bandeaux de forces discrets correspondants, modélisant 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 particulière θ. Les $d\mathcal{FB}$ sont représentés sous la forme de cartes de chaleur.

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 [20, 26] pour l'histogramme de forces. 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 que ce soit la même sur les deux objets [16].

3.4 Reconnaissance de relations spatiales

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 [32], ou utiliser des règles d'évaluation prédéfinies à partir d'analyses théoriques comme dans [19]. Nous nous proposons d'utiliser l'apprentissage pour traduire notre descripteur en relations spatiales en langage naturel, comme cela est fait dans [32] 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.

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).

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 fonction de transformation 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 directement en comparant ces primitives si nous recherchons 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. Cette extension fait partie des perspectives de notre travail.

Chaîne globale L'approche proposée a l'avantage d'être adaptée à tout type d'objets et à toute configuration avec peu de données d'apprentissage, en utilisant le $d\mathcal{FB}$ comme représentation intermédiaire. Cependant, il nécessite de travailler avec des objets binaires pour calculer le $d\mathcal{FB}$, donc une étape de segmentation est nécessaire pour

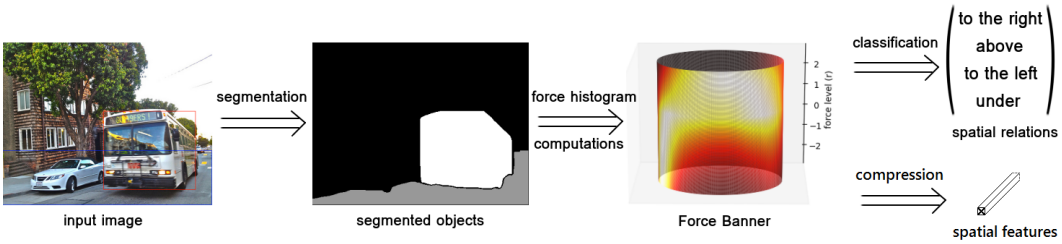

FIGURE 4 – Chaîne globale sur une image naturelle (bandeau de forces représenté sous la forme d'un cylindre $3D$).

pouvoir utiliser des images naturelles. Ainsi, 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 4) : (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 4). Nous avons commencé par considérer quatre relations spatiales simples, mais cela peut être étendu à davantage de relations si des annotations sont disponibles.

4 Étude expérimentale

4.1 Données

Dans cette étude expérimentale, deux types de données ont été utilisées : (1) des données synthétiques : 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 ; (2) des données réelles : il s'agit d'images naturelles dont les objets sont extraits par une étape de segmentation. Pour cela nous avons utilisé une image de télédétection déjà segmentée, que nous avons découpée en tuiles. Les données synthétiques ont été utilisées pour entraîner et tester le modèle, tandis que les données réelles ont été utilisées uniquement pour le test.

Nous avons ainsi obtenu les trois jeux de données suivants :

- *SimpleShapes1* (S1) : 1 000 images de paires d'objets obtenues à partir de 10 formes différentes représentant des objets simples (maisons, avions, voitures, etc.). Toutes ces paires ont en général des configurations simples puisque tous les objets sont totalement remplis et ont des tailles proches ;
- *SimpleShapes2* (S2) : 1 280 images de paires d'objets obtenues à partir de 8 formes géométriques simples, petites ou allongées. Ces paires ont des configurations plus complexes, avec des objets de tailles variables ;
- *image SIG* (SIG) : 211 images contenant des paires d'objets extraits d'une image de télédétection (maisons, routes, rivière, champs). Il est intéressant de noter qu'il contient des formes constituées de plusieurs parties non connexes.

Toutes ces images sont de dimension 224×224. Elles ont été annotées manuellement pour fournir une relation spatiale pour chaque scène selon 4 classes directionnelles

TABLE 1 – Composition des différents jeux de données en termes de complexité, des relations faciles ($N1$) aux relations totalement ambiguës ($N4$, exclues des tests).

	Total	$N1$	$N2$	$N3$	$N4$
SimpleShapes1 (S1)	1000	63.4%	15.5%	7.7%	13.4%
SimpleShapes2 (S2)	1280	61.5%	17.0%	9.5%	12.0%
image SIG (SIG)	211	59.7%	23.2%	7.1%	10.0%

("*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 3).

En raison du caractère aléatoire du processus génératif, les ensembles 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, 1 127 dans S2 et 190 dans l'image SIG. La composition de chaque jeu de données est donnée dans le Tableau 1, tandis que la Figure 5 donne un exemple pour chaque relation et chaque niveau de complexité.

4.2 Protocole expérimental

Modèle retenu Comme introduit dans la Section 3.4, nous suggérons d'utiliser un CNN $2D$ pour manipuler les bandeaux de forces. 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 28×28 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 5×5 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 224×224 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 224×245).
Nous avons tout de même utilisé un réseau *SqueezeNet* pré-entraîné sur IMAGENET à des fins de comparaison, et nous avons obtenu des résultats très proches.

Méthodes comparatives Afin de mesurer l'intérêt de notre approche, nous la comparons à plusieurs références :

- *bbox coords* : classification supervisée basée sur les coordonnées des rectangles englobants des objets. Un perceptron multi-couches (MLP) avec activation ReLU est utilisé pour cela, avec 4 couches et 38 501 paramètres.
- *bbox image* : classification supervisée basée sur l'image binaire, en entraînant un CNN similaire à celui utilisé pour le $d\mathcal{FB}$. Comme pré-traitement pour aider à la classification, l'image est réduite au rectangle englobant contenant les deux objets et mise à l'échelle 224×224. Cette méthode est proche de celle proposée dans [15], mais dans notre approche les objets sont segmentés et leurs classes ne sont pas considérées.
- *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.

Pour terminer, nous comparons également à la fusion des sorties des deux modèles entraînés sur les images binaires et leurs bandeaux (méthode $d\mathcal{FB}$*+image*), 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.

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 $0,025$. La Figure 3 présente des exemples des bandeaux ainsi obtenus pour les différents jeux de données. Sur la Figure 5, les images et leurs bandeaux de forces sont rangés selon la relation spatiale et le niveau de complexité de cette relation.

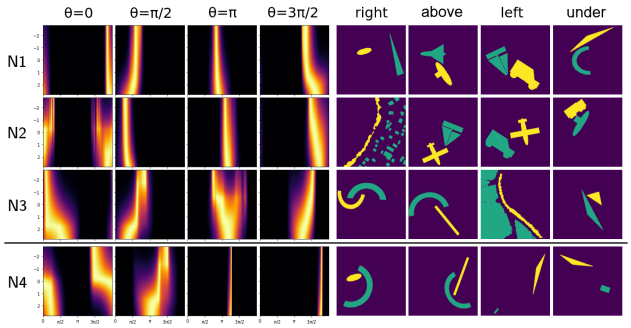

FIGURE 5 – Exemples de bandeaux de forces pour les différentes classes (colonnes) et les différents niveaux d'ambiguïté (lignes), de $N1$ (facile) à $N4$ (non décidable).

Entraînement des modèles Les modèles sont entraînés sur la fusion des jeux de données S1 et S2, le jeu SIG étant réservé au test. Afin d'éviter un biais d'apprentissage dans l'évaluation tout en utilisant l'ensemble des données pour l'entraînement et la validation (prédiction sur une autre partie du même jeu de données), 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% restant). Les mêmes proportions de chaque classe et chaque niveau de difficulté sont conservées dans chaque sous-ensemble afin d'avoir une performance comparable. Tous les modèles sont entraînés avec une fonction d'entropie croisée comme fonction de coût, un optimiseur *Adam*, un taux d'entraînement de 10^{-4} réduit 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 $\epsilon = 10^{-8}$).

4.3 Résultats et discussion

Résultats préliminaires Des exemples illustratifs de bandeaux de force obtenus pour les différentes relations spatiales et les différents niveaux de difficulté sont donnés sur la Figure 5. On peut voir 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.

TABLE 2 – Résultats de classification (taux de bonne classification maximal sur 50 epochs – OA en %, et écart-type sur 4 partitions de validation croisée – STD en %) sur les ensembles de test pour les différentes méthodes.

	dF-banner		*bbox image*		*dFB+image*		*bbox coords*		*quadrants*	
Test data	**OA**	**STD**	**OA**	**STD**	**OA**	**STD**	**OA**	**STD**	*centers*	*masks*
S1+S2	**97,89**	0,53	93,08	0,87	97,24	1,15	96,54	0,19	95,48	95,08
S1	**98,96**	0,45	96,33	1,43	98,61	0,55	97,92	0,30	95,84	95,96
S2	**97,32**	1,01	91,72	1,79	96,96	1,64	95,83	0,58	95,21	94,41
S1+S2_N2-3	97,75	0,41	95,49	1,71	**98,57**	1,03	93,03	0,82	87,41	86,54
S1+S2_N3	95,83	0,00	91,67	2,95	**96,88**	2,69	86,98	1,04	81,41	82,91
SIG	**99,47**	0,00	87,37	1,36	97,63	0,52	96,19	0,26	96,84	94,74

Étude comparative Le Tableau 2 présente les résultats de classification obtenus en utilisant différents sous-ensembles du jeu de données. 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.

Tout d'abord, on peut observer que la tâche de classification 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. 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 $0,1\%$ et produit même un sur-entraînement pour certaines méthodes.

Les meilleurs scores sont obtenus par la méthode $d\mathcal{FB}$ ou par sa fusion avec *bbox 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. 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 $d\mathcal{FB}$, 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 , on vérifie que la performance reste stable. On peut alors observer que l'écart entre le $d\mathcal{FB}$ (ou sa fusion avec *bbox image*) et les autres méthodes augmente, puisque le score pour le $d\mathcal{FB}$ reste élevé alors que les méthodes comparatives ont beaucoup plus de difficultés à appréhender les configurations plus complexes. Enfin, on constate que la méthode $d\mathcal{FB}$+*image* dépasse le $d\mathcal{FB}$ seul dans ces tests, donc chaque approche apporte du contenu complémentaire dans ces cas difficiles.

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.

5 Conclusion

Nous avons introduit le bandeau de forces comme descripteur de position relative entre objets binaires. Cette extension de l'histogramme de forces [23] a plusieurs avantages : 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 ; 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 ; il présente l'information spatiale de manière claire et précise ; 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 ; 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 être facilement 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. 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.

Crédits

Ce travail mené au LIPADE, est financé par Magellium, avec le soutien de l'Agence de l'Innovation de Défense.

Références

[1] J. F. Allen. Maintaining knowledge about temporal intervals. *Comm. of the ACM*, 26(11) :832–843, 1983.

[2] I. Bloch. Fuzzy Relative Position between Objects in Image Processing : A Morphological Approach. *IEEE TPAMI*, 21(7) :657–664, 1999.

[3] Isabelle Bloch. Fuzzy spatial relationships for image processing and interpretation : A review. *IVC*, 23(2) :89–110, 2005.

[4] Isabelle Bloch, Olivier Colliot, and Roberto M. Cesar. On the Ternary Spatial Relation "Between". *IEEE TSMC*, 36(2) :312–327, 2006.

[5] Andrew R Buck, James M Keller, and Marjorie Skubic. A memetic algorithm for matching spatial configurations with the histograms of forces. *IEEE TEC*, 17(4) :588–604, 2013.

[6] Roberto M. Cesar, Endika Bengoetxea, and Isabelle Bloch. Inexact graph matching using stochastic optimization techniques for facial feature recognition. In *ICPR*, pages 465–468, 2002.

[7] M. Clément, M. Garnier, C. Kurtz, and L. Wendling. Color Object Recognition based on Spatial Relations between Image Layers. In *VISAPP*, pages 427–434, 2015.

[8] M. Clément, C. Kurtz, and L. Wendling. Bags of Spatial Relations and Shapes Features for Structural Object Description. In *ICPR*, pages 1995–2000, 2016.

[9] M. Clément, C. Kurtz, and L. Wendling. Learning spatial relations and shapes for structural object description and scene recognition. *PR*, 84 :197–210, 2018.

[10] Michaël Clément, Adrien Poulenard, Camille Kurtz, and Laurent Wendling. Directional Enlacement Histograms for the Description of Complex Spatial Configurations between Objects. *IEEE TPAMI*, 39(12) :2366–2380, 2017.

[11] A. G. Cohn, B. Bennett, J. Gooday, and N. M. Gotts. Qualitative spatial representation and reasoning with the region connection calculus. *GeoInformatica*, 1(3) :275–316, 1997.

[12] Olivier Colliot, Oscar Camara, and Isabelle Bloch. Integration of fuzzy spatial relations in deformable models – Application to brain MRI segmentation. *PR*, 39(8) :1401–1414, 2006.

[13] Adrien Delaye and Eric Anquetil. Learning of fuzzy spatial relations between handwritten patterns. *IJDMMM*, 6(2) :127–147, 2014.

[14] J. Freeman. The Modelling of Spatial Relations. *CGIP*, 4(2) :156–171, 1975.

[15] M. Haldekar, A. Ganesan, and T. Oates. Identifying spatial relations in images using convolutional neural networks. In *IJCNN*, pages 3593–3600, 2017.

[16] M. Jazouli, J. Wadsworth, and Pascal Matsakis. Normalization of the Histogram of Forces. In *ICPRAM*, pages 630–639, 2019.

[17] L. T. Kóczy. On the description of relative position of fuzzy patterns. *PRL*, 8(1) :21–28, 1988.

[18] Nicolas Loménie and Daniel Racoceanu. Point set morphological filtering and semantic spatial configuration modeling : Application to microscopic image and bio-structure analysis. *PR*, 45(8) :2894–2911, 2012.

[19] Pascal Matsakis, J M Keller, Laurent Wendling, Jonathan Marjamaa, and Ozy Sjahputera. Linguistic description of relative positions in images. *IEEE TSMC*, 31(4) :573–88, 2001.

[20] Pascal Matsakis, James M. Keller, Ozy Sjahputera, and Jonathon Marjamaa. The use of force histograms for affine-invariant relative position description. *IEEE TPAMI*, 26(1) :1–18, 2004.

[21] Pascal Matsakis and Mohammad Naeem. Fuzzy Models of Topological Relationships Based on the PHI-Descriptor. In *FUZZ-IEEE*, pages 1096–1104, 2016.

[22] Pascal Matsakis, Mohammad Naeem, and Farhad Rahbarnia. Introducing the Φ-Descriptor – A Most Versatile Relative Position Descriptor. In *ICPRAM*, pages 87–98, 2015.

[23] Pascal Matsakis and Laurent Wendling. A New Way to Represent the Relative Position between Areal Objects. *IEEE TPAMI*, 21(7) :634–643, 1999.

[24] K. Miyajima and A. Ralescu. Spatial organization in 2D segmented images : Representation and recognition of primitive spatial relations. *FSS*, 65(2) :225–236, 1994.

[25] Mohammad Naeem and Pascal Matsakis. Relative Position Descriptors – A Review. In *ICPRAM*, pages 286–295, 2015.

[26] Jingbo Ni and Pascal Matsakis. An equivalent definition of the histogram of forces : Theoretical and algorithmic implications. *PR*, 43(4) :1607–1617, 2010.

[27] D. A. Randell, Z. Cui, and A. G. Cohn. A spatial logic based on regions and connection. In *KR*, pages 165–176, 1992.

[28] A. Rosenfeld and R. Klette. Degree of adjacency or surroundedness. *PR*, 18(2) :169–177, 1985.

[29] KC Santosh, Laurent Wendling, and Bart Lamiroy. BoR : Bag-of-relations for symbol retrieval. *IJPRAI*, 28(06) :1450017, 2014.

[30] Salvatore Tabbone and Laurent Wendling. Color and grey level object retrieval using a 3D representation of force histogram. *IVC*, 21(6) :483–495, 2003.

[31] Maria Carolina Vanegas, Isabelle Bloch, and Jordi Inglada. A fuzzy definition of the spatial relation “surround” - Application to complex shapes. In *EUSFLAT*, pages 844–851, 2011.

[32] Xiaomei Wang and J. M. Keller. Human-based spatial relationship generalization through neural/fuzzy approaches. *FSS*, 101(1) :5–20, 1999.