Force Banner for the recognition of spatial relations

Robin Deléarde, Camille Kurtz, Philippe Dejean, Laurent Wendling

Force Banner for the recognition of spatial relations

Robin Deléarde*†, Camille Kurtz*, Philippe Dejean† and Laurent Wendling*
*Université de Paris, LIPADE EA 2517, Paris, France
†Magellium, Toulouse, France
{firstname.lastname}@{u-paris,magellium}.fr

***Abstract*—Studying the spatial organization of objects in images is fundamental to increase both the understanding of a sensed scene and the explainability of the perceived similarity between images. This leads to the fundamental problem of handling spatial relations: given two objects depicted in an image, or two parts in an object, how to extract and describe efficiently their spatial configuration? Dedicated descriptors already exist for this task, like the efficient force histogram. In this article, we introduce the *Force Banner*, which extends it to two dimensions by using a panel of forces (attraction and repulsion), so as to benefit from more expressiveness and to model rich spatial information. This descriptor can be used as an intermediate representation of the image dedicated to the spatial configuration, and feed a classical $2D$ Convolutional Neural Network (CNN) to benefit from their powerful performances. As an illustration of this, we used it to solve a classification problem aiming to discriminate simple spatial relations, but with variable configuration complexities. Experimental results obtained on datasets of images with various shapes highlight the interest of this approach, in particular for complex spatial configurations.**

I. Introduction

In recent years, taking spatial relationships into account in image analysis processes has been a hot topic studied by the computer vision community, and more generally in the pattern recognition domain. In fact, it can be stated that the spatial organizations between image components are fundamental in the human perception of image understanding. Therefore, the spatial relations between the regions composing a scene can be considered as important features, to recognize the nature of the scene itself for instance. However, as far as we know they are seldom used for image recognition, mostly because they often suffer from strong structural constraint issues. Thus, most of the state-of-the-art methods dedicated to the recognition of complex scenes usually rely on a structural or statistical description of the image content, summarizing different image features such as outer contour, geometry or texture and color effects. As a limit, these different types of imaging features are sometimes not discriminant enough to successfully describe image contents composed of objects mutually arranging with complex spatial configurations.

In order to be able to integrate this type of spatial information in future recognition systems, a preliminary question that can be asked may be the following: is it possible to automatically recognize a spatial relationship between a pair of objects present in the content of a scene? In this article, we consider this as a classification problem. Given two objects and their relative position in the image, we want to build a model being able to predict the spatial relation characterizing their spatial arrangement (*e.g.*, object A is "*to the left of*" object B).

In parallel, deep learning based strategies (such as Convolutional Neural Networks (CNNs)) have been proposed in the computer vision community to efficiently exploit the discriminative aspects of local features in images for various tasks. Such models have led to outstanding results in image classification tasks, but one of their inherent downside is precisely their weak ability to take into account spatial information, because images are represented as orderless collections of local features. Furthermore, they are difficult to exploit to directly recognize spatial relationships (that are often ambiguous) between objects because the convolutional features, computed from the initial image space, carry a point of view that may be too local.

In this article, we propose to combine the advantages of traditional approaches a.k.a. relative position descriptors to those of CNNs to answer the problem of the recognition of spatial relations. Rather than trying to learn a spatial relationship directly from the initial image space, as it is the case with standard CNNs based approaches, we propose in Section III an intermediate representation of the image, capturing information of relative positions between a couple of objects, and we train the CNNs to recognize the spatial relationship from this rich representation. Such new representation (called Force Banner) extends the concept of the Force Histogram [1] since it captures the relative position between objects using a panel of forces, from attraction to repulsion. Like the Force Histogram, it takes into account the structural shapes of the objects and their distance in a directional manner.

The remainder of this article is organized as it follows. Section II reviews some related works in the context of this paper. Section III presents our methodological contribution. In Section IV, we propose an experimental study, where a dataset with variable complexity in object configurations is considered to illustrate the interest of our approach. We show that the combination of the Force Banner with the original image features allows to better recognize complex spatial relations. Finally, a conclusion that emphasizes the perspectives of this work is made in Section V.

II. Related works

Many studies have been conducted for the analysis of spatial relations in different application domains of pattern recognition and computer vision, with the common objective of describing the spatial arrangement of objects in images [2].

We can distinguish in the literature two main research axes based on strong dual concepts [3]: the concept of spatial relation, and the one of relative position of an object with regards to another.

In the first axis, a spatial relation such as "*to the left of*" is considered, and a fuzzy evaluation of this relation is obtained for two given objects. For instance, the fuzzy landscape framework [4] focuses on this type of evaluations. This approach is based on a fuzzy modeling of spatial relations directly in the image space, using morphological operations. Typical applications include graph-based face recognition [5], brain segmentation from MRI [6], or handwritten text recognition [7].

In the second axis, the relative position of an object with regards to another one can have a representation of its own, from which it is possible to derive evaluations of spatial relations. Different spatial relations can be assessed from this intermediate representation and the associated descriptors can be integrated into pattern recognition processes to match similar spatial configurations or to predict spatial relations. A typical relative position descriptor is the Force Histogram [1], which is a generalization of the Angle Histogram [8]. Notably, Force Histograms are isotropic and less sensible to discretization issues, while also allowing to explicitly take into account the distance between objects, depending on the application needs. Force Histograms are involved in several application domains such as linguistic descriptions [9], [10], scene matching [11] or content-based image retrieval [12], [13].

Note that many other approaches were also proposed for modeling more specific spatial relations such as the "*surrounded by*" relation [14], the "*between*" relation [15], [16], or even the "*enlaced by*" relation [17]. Other recent works introduced the ϕ-descriptor [18], [19], which provides a generic framework to assess any spatial relation from a set of specific operators, based partially on Allen intervals [20]. This descriptor provides an important advancement, while requiring an extraction of a set of suitable operators dedicated to each usual spatial relation.

Going back to the model of Force Histograms, the authors of [21], [22] introduced the Force Histogram Decomposition (FHD), a graph-based hierarchical descriptor that allows to characterize the spatial relations and shape information between the pairwise structural subparts of objects. A novel "bags-of-relations" framework based on such descriptors is used to produce discriminative structural features that are suited for particular object classification tasks. An advantage of this learning procedure is its compatibility with traditional bags-of-features frameworks, allowing for hybrid representations that gather structural and local features. The authors of [22] also shown the importance and the complementarity of the different forces (negatives for repulsion, positives for attraction) involved in Force Histograms to improve the recognition of complex classes.

More specifically, the recognition of spatial relations (given two objects in an image, what is their spatial relation?) can also be considered as a classification task. In this context, this problem (which has long remained a problem of pattern recognition) has been covered by the field of computer vision, notably with the arrival of deep learning architectures. The task often consists in visual relationship detection: given an image, the algorithm predicts "subject-predicate-object" triplets as well as the bounding boxes of the objects. In contrast, our task is classification rather than detection: the object pairs are given (for instance as a raster binary image), and we aim at a finner evaluation of relation understanding. Most of the approaches from the state-of-the-art [23]–[25] rely on CNNs architectures where a model is trained from the coordinates of the bounding boxes of the objects (visual features), and potentially their semantics (*e.g.*, chairs, guitars) called language features, to predict the spatial relations.

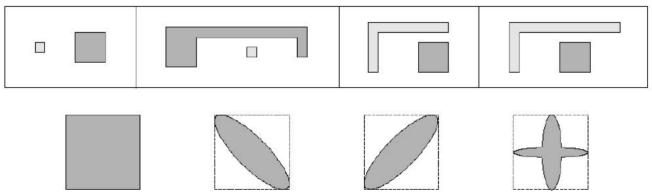

Figure 1. Centroid and bounding box spatial location ambiguities.

Such approaches require large amount of annotated data to get accurate results. To get such collections, we can cite crowdsourcing initiatives like Open Images, Visual Genome [26], or SpatialSense [27], that led to the creation of datasets containing a significant number of visual relations. SpatialSense [27] is for example constructed through adversarial crowd-sourcing, in which human annotators are asked to find spatial relations that are difficult to predict using simple clues such as bounding boxes or common configurations. Such annotated datasets may allow for proper benchmarking of spatial relation recognition techniques. However, they often contain only the bounding boxes of the objects of interest present in the image content, since most of the state-of-the-art approaches in computer vision rely on these features for spatial relation recognition.

Old but often overlooked work has shown that standard "all or nothing" mathematical relationships are clearly not suitable, and Freeman [28] (in 1975) suggested to use fuzzy relationship. However, computers have not been able to effectively model these vital spatial concepts. For instance, many authors assimilated 2D objects to very elementary entities such as a point (centroid) or a (bounding) rectangle. The procedure is practical and convenient in most of the cases, but cannot be hoped to give a satisfactory modeling, as pointed out by Rosenfeld [29] in 1985 (see Figure 1).

In this work, on the one hand, in order to predict a spatial relation, our motivations are to evaluate the relative position of a 2D object A compared to another object B by a set of functions, called Force Banner, corresponding to either repulsion or attraction forces between the two objects. The object A is the argument, and the object B the referent. For

any direction θ, the weight value of the arguments that can be found in order to support the proposition "A is in direction θ of B" is calculated from several force representations.

On the other hand, our objective is to study the possibility of a CNN to apprehend these spatial relationships by simply considering a set of training sample data representing the four main cardinal directions between two objects. In other words, the goal is to attest the potential of a CNN to learn the description of directional relations, by considering a single object representing the argument and another one the referent.

III. Proposed approach for the recognition of spatial relations

We describe hereinafter the proposed approach for spatial relation recognition. We propose a novel representation, called Force Banner, derived from the concept of Force Histogram [1] (a reminder on this one is given in Section III-A), modeling directional information about the relative position of pairs of objects composing a scene. Such a representation is described in Section III-B. Given a binary image containing a pair of objects, it captures the relative position between objects using a panel of forces (attractives to repulsives), that take into account the structural shapes of the objects and their distance. Force Banners are used to feed a classical $2D$ CNN for the recognition of spatial relations, benefiting from pre-trained models and fine-tuning (Section III-C).

A. Notions on Force Histograms [1]

The Force Histogram model was initially introduced in [1]. In this section, we briefly recall the main definitions and principles of this model, which constitutes the basis of our novel approach.

Force Histograms (thereafter noted F-Histograms) aim to evaluate and characterize the directional spatial relations between binary objects in images. The model relies on the definition of a force of attraction between points. Given two points located at a distance d from each other, their force of attraction is as follows:

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

where r characterizes the kind of force processed. Instead of directly studying all pairs of points between the two objects, the force of attraction between two one-dimensional segments is considered. Let I and J be two segments on a line of angle θ, D^θ_{IJ} the distance between them and $|.|$ the segment length. The force of attraction f_r of segment I with regards to segment J is given by:

$$f_r(I,J) = \int_{D^\theta_{IJ}+|J|}^{|I|+D^\theta_{IJ}+|J|} \int_0^{|J|} \varphi_r(u-v)\,\mathrm{d}v\,\mathrm{d}u. \tag{2}$$

Given two binary objects A and B, a θ-oriented line in the image forms two sets of segments belonging to each object: $\mathcal{C}_A = \cup\{I_i\}_{i=1..n}$ and $\mathcal{C}_B = \cup\{J_j\}_{j=1..m}$. The mutual attraction between these segments is defined as:

$$F_r(\theta, \mathcal{C}_A, \mathcal{C}_B) = \sum_{I\in\mathcal{C}_A} \sum_{J\in\mathcal{C}_B} f_r(I,J). \tag{3}$$

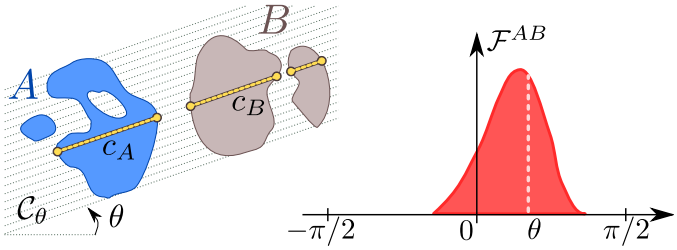

Figure 2. Illustration of the F-Histogram computation scheme. The force of attraction between A and B along the direction θ is the integral sum of forces computed on longitudinal cuts (C_A, C_B) [1].

Then, the set of all θ-oriented parallel lines $\mathcal{C}_\theta$ going through the whole image gives us the global attraction $F_r^{AB}(\theta)$ between A and B along a direction θ. Figure 2 summarizes the process for a given direction. Finally, the F-Histogram $\mathcal{F}_r^{AB}$ is obtained by computing F_r^{AB} onto a set of angles $\theta \in [0, 2\pi[$, summarizing the relative position of a binary object A (commonly called the argument) with regards to a binary object B (the referent) in a circular way.

B. Towards the Force Banner

Usually two levels of forces are widely used in the literature to assess spatial relation between a couple of objects:

- $r = 0$ relies on constant forces which are independent of the distance between objects. In some extent this approach is based on the handling of an isotropic histogram of angles;
- $r = 2$ relies on gravitational forces where more importance is given to closer points.

Evaluating F_0^{AB} gives an overview of the scene, but it is often too cautious. Such behavior can be corrected by considering F_2^{AB}, which focuses on close-up views between the object A and B. However, a complex situation can give a contradictory opinion (sometimes excessively pessimistic and sometimes excessively optimistic). In [10] it was shown that the combination of these two types of forces can provide an efficient and robust system for obtaining a linguistic description of a scene. And in [22] it was shown that negative forces ($r = -2$ for instance) embedded in a bag of relations can also bring another point of view during a classification process. Thus, the description potential of r can be different depending on the complexity of the scenes considered.

Our idea is to provide in a single representation, called Force Banner, a series of $\mathcal{F}_r^{AB}$, to better take into account the complex description of a situation, and to use it as input of a CNN (see section III-C) to extract automatically the most discriminative features (forces versus directions), providing a convenient model for a supervised classification task and benefiting from CNN good performance.

Let A and B be two objects and let r and θ be two real numbers such that $r \in [r_s, r_e]$ and $\theta \in [0, 2\pi[$. The Force Banner FB^{AB} is defined as it follows:

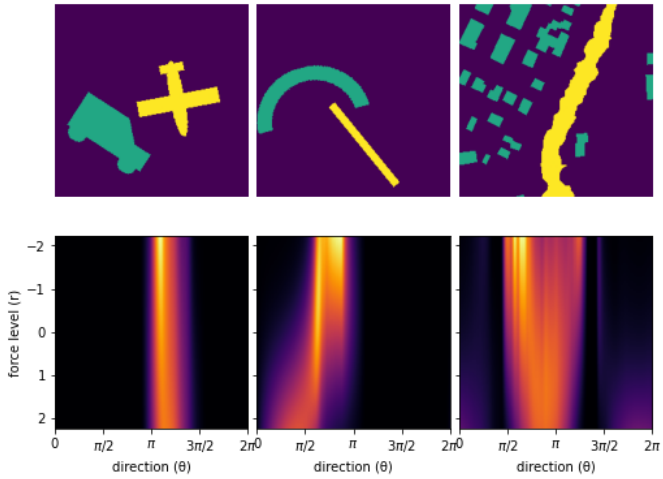

Figure 3. (First line) Illustrative samples of the considered dataset of binary shapes: pair of simple objects, pair of geometrical shapes, pair of GIS objects (houses and road). The referent is in yellow while the argument is in green. (Second line) Corresponding discrete Force Banners modeling the relative position between the argument and the referent objects. Each row corresponds to a particular force r while each column represents the force in a particular direction θ. Note that for visualization purpose, the discrete Force Banners are represented here as heatmaps.

$$FB^{AB} : [0, 2\pi[\times [r_s, r_e] \rightarrow \mathbb{R}_+ \\ (\theta, r) \mapsto F_r^{AB}(\theta) \tag{4}$$

Considering different values of r, the assessable property of the Force Banner depends on the location of A and B. If the objects A and B are both disjoint and non tangent then FB^{AB} is assessable for any value of r. Otherwise if $[r_s, r_e] \subset]-\infty, 2[$ then disjoint and tangent object can be considered to make FB^{AB} assessable. If $[r_s, r_e] \subset]-\infty, 1[$ then any couple of objects can be considered to make FB^{AB} assessable.

Furthermore it is easy to show from [1] that the Force Banner is invariant with regards to translations and scaling transformations (after normalization). It is also isotropic if a circular shift along the θ-axis is performed to take into account the effects of the rotation.

C. Chosen CNN model

Convolutional Neural Networks refer to a family of deep learning algorithms. Susch systems are composed of two parts. The first one is designed to feature extraction, it has many neuron layers that compute the convolutions of the previous ones. The neurons of each layer are activated by non-linear functions (*e.g.*, sigmoid, ReLU) in order to keep the most representative features (high order features). We find also max-pooling layers between convolutional layers to reduce faster the size of the intermediate features and the number of parameters to be computed to define the network, and hence to control over-fitting. The second part is the classifier, using high order features to make the decision. Generally, it is a fully connected layer that provides a probability vector, on which is plugged a softmax function to predict the class label of input data.

We have chosen the SqueezeNet model [30] but any other $2D$ CNN model can be used. SqueezeNet has interesting properties, like the same accuracy level as the AlexNet model on the ImageNet dataset with much fewer parameters, which makes the training faster. The architecture of SqueezeNet introduces a new module called Fire composed of a squeeze layer using 1×1 convolution filters followed by expand layer that contains a mix of 1×1 and 3×3 convolution filters. Also, its classifier is based on a global average pooling over feature maps, potentially decreasing the overfitting effect. Its global architecture is illustrated on Figure 4.

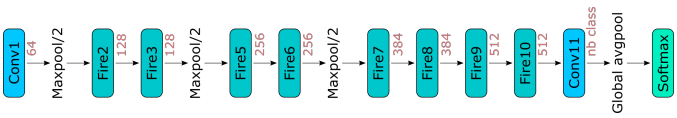

Figure 4. SqueezeNet architecture [30]

We used the PyTorch implementation of SqueezeNet, pre-trained on ImageNet. This pre-training provides a model which is already quite good for any classification task on images, after replacing the last layer by another one dedicated to the considered task and training only that layer, which is called transfer learning. However, numerous studies showed the interest of post-training the whole model on the proper data, especially here where the data are not real images but a stack of histograms. This step called fine-tuning allows to specialize the network to an ad-hoc task, leading ultimately to better performances. We then considered this approach in our experiments.

IV. Experimental study

To evaluate the interest of our approach, the proposed Force Banner representation is involved in a spatial relation recognition task on a dataset containing various configurations (from simple to more complex ones) of object pairs. The objective is also to show its complementarity with the original image representation, when considering a CNN-based learning and classification strategy, here on a classification task with 4 classes (North, South, East, West). This protocol is similar to the one used in [9] to evaluate the histogram of angles, which was based on a simpler neural network architecture and much fewer images.

A. Data

As mentioned in Section II, recent initiatives in the computer vision domain are currently leading to the production of large databases dedicated to spatial relation recognition [27]. As a limit, such datasets only contain the bounding boxes of the objects of interest present in the image content, or they do not provide together the spatial relationships and the segmentation of the involved objects. Since our hypothesis is that considering the shapes and the structures of the objects should be investigated to assess their configuration, we need a

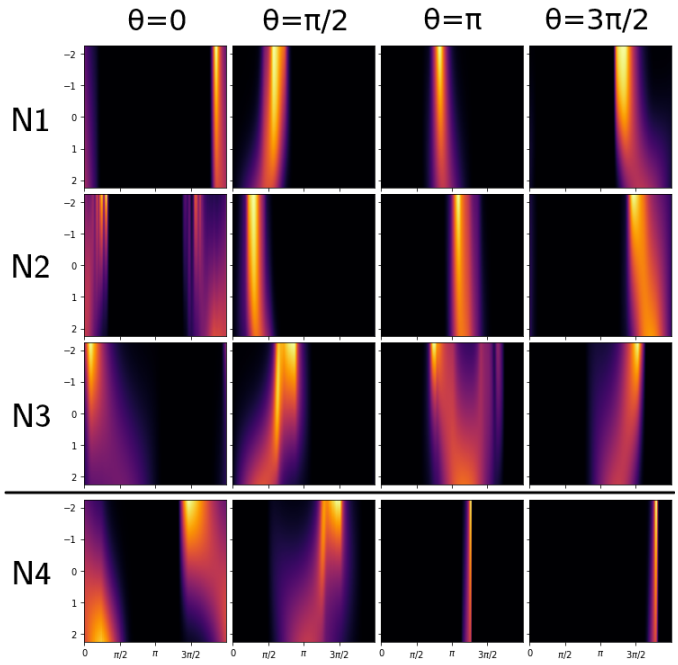

Figure 5. Samples of discrete Force Banners for the different classes (columns) and for different ambiguity levels (rows), from level N1 (easy) to level N4 (not decidable). For each class a privileged direction can be deduced, i.e. a small range in abscissa, which is clear for the simple cases and gets fuzzy or too far from a privileged direction as the ambiguity increases.

delimitation of these objects, so we built a new image database more suitable for our experiment.

To this end, we randomly generated multiple combinations of binary shapes that we drew in an image on a homogeneous background. Such idea was already considered in various articles from the state-of-the-art in description of spatial relations [4], [19]. The position of the objects in the image is also randomly decided to guarantee a greater diversity. We have only avoided that the binary shapes are strictly disjoint to guarantee the assessable property of the Force Banner. The image size is 224×224 to be compatible with the input of networks. Technically, the produced images are gray-level images where each object (referent versus argument) appears with a specific gray-level, and the background in black.

Two families of shapes were considered:

1) shapes corresponding to geometric shapes (triangle, rectangle, ellipse) and simple objects made from those shapes (houses, planes, cars, etc., in different views);
2) shapes corresponding to urban objects extracted from a remote sensing image (houses, roads, river, agricultural crop-fields).

This has led to the creation of two datasets subsequently named *SimpleShapes* (2280 images) and *GIS* (211 images). Figure 3 presents some samples of the dataset (in color for visualization purpose). The second dataset will be used as a realistic test case in our experiments. Moreover, it is interesting to note that, contrary to the first one, it contains shapes made of several parts that are not connected, which may result in more complex situations.

These images were then manually annotated by three experts to provide a spatial relationship for each scene. For each image, one object was considered as the referent and one as the argument, and a spatial relation is chosen. We consider the opposite relation as symmetric. For complex and ambiguous cases between experts, a vote is taken.

Four classes were considered during the annotation phase (North, South, East, West). Due to the randomness of the generative process, the datasets contain various spatial configurations ranging from simple configurations to more complex ones that can lead to more ambiguous spatial situations. The images were also sorted according to the complexity and ambiguity level of the spatial relation (in 4 different levels N1–N4), so as to evaluate separately on each part. Really ambiguous cases that were not decidable (N4) were rejected from the datasets in the experiments, leading to 1953 images in *SimpleShapes* and 190 images in *GIS*.

B. Experimental protocol

We provide hereinafter the experimental protocol followed in this applicative study.

1) Discrete Force Banner: Using raster data, a matrix $\widetilde{dFB}^{AB}$ is obtained from a discrete approximation of the Force Banner FB^{AB}. Let us consider $\Theta = \{\theta_1, \theta_2, \ldots, \theta_{|\Theta|}\}$ a set of consecutive directions defined from a constant step $\delta_\theta \in \mathbb{R}$ (that is $\theta_{i+1} = \theta_i + \delta_\theta$ and $\theta_0 = 0$ and $\theta_{|\Theta|} = 2\pi - \delta_\theta$). Let $r_s \in \mathbb{R}$ and $r_e \in \mathbb{R}$ be two forces and a set of forces between theses two bounds $R = \{r_s, r_s + \delta_r, \ldots, r_e\}$ with δ_r a force discretization step. Each row of the matrix is normalized by its own area to ensure the same importance for each directional relation. Then, $\widetilde{dFB}^{AB}$ is defined as follows:

$$\widetilde{dFB}^{AB} = \begin{vmatrix} \mu_{\theta_0,r_s} & \cdots & \mu_{\theta_{0+i\delta_\theta},r_s} & \cdots & \mu_{\theta_{2\pi-\delta_\theta},r_s} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \mu_{\theta_0,r_j} & \cdots & \mu_{\theta_{0+i\delta_\theta},r_j} & \cdots & \mu_{\theta_{2\pi-\delta_\theta},r_j} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \mu_{\theta_0,r_e} & \cdots & \mu_{\theta_{0+i\delta_\theta},r_e} & \cdots & \mu_{\theta_{2\pi-\delta_\theta},r_e} \end{vmatrix} \quad (5)$$

and

$$\mu_{\theta_i,r_j} = F^{AB}_{r_j}(\theta_i)/||F^{AB}_{r_j}(.)|| \quad (6)$$

with $r_j = r_{s+j\delta_r}$.

This means that all the values are in the range $[0, 1]$ to take into account the scale factor and the forces taken independently have the same relevance. To be compatible with the CNN input data without having to rescale images, we considered in these experiments 224 different directions ($|\Theta| = 224$ and $\delta_\theta = 2\pi/224$), and 224 forces from $r_s = -2.24$ to $r_e = 2.24$, with a step of 0.02.

The discrete Force Banner $\widetilde{dFB}^{AB}$ can then be encoded as a $2D$ gray-scale image of size 224×224 where each row corresponds to a particular force r while each column represents the force in a particular direction θ. Figure 3 and

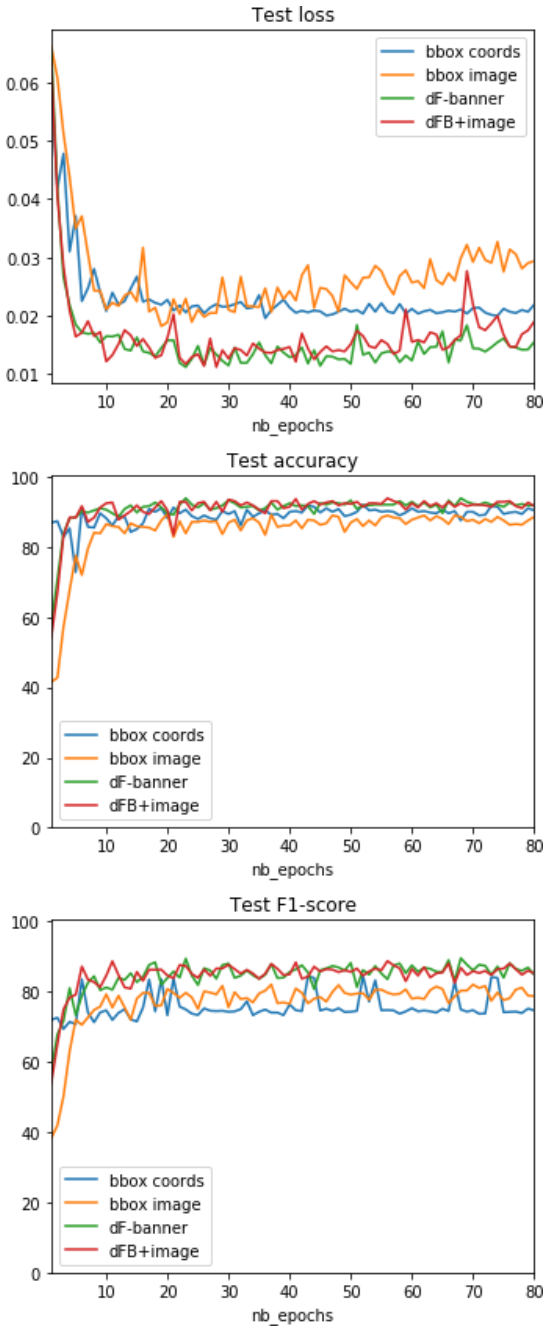

Figure 6. Test/validation loss and accuracy curves obtained on the *SimpleShapes* dataset (Train & Test on N1+N2+N3), when the models are trained with the different data representations with the same splits.

Figure 5 present some samples of discrete Force Banners (shown as heatmaps only for visualization purpose).

2) Studying Force Banner vs. original image information: To quantitatively show that the recognition of the spatial relation between a pair of objects may be enhanced by considering a discrete Force Banner instead of directly considering the original image representation, we proceed to the following experiment.

To classify images from the datasets, we consider as input of the CNN network either their discrete Force Banner representations (*dF-banner*), or the binary images cropped to the bounding box containing the two objects and rescaled to the initial size 224×224 (*bbox image*), or a combination of the two since they carry additional information (*dFB + image*). Cropping the raw image is applied as an intuitive pre-treatment to help the CNN. Doing so, we have a method that is quite naive but interesting enough to be used as a comparison. To combine the two representations, the two CNN models dedicated to the cropped image and to its discrete Force Banner are trained jointly on the same indices, and the fusion is performed with an additional fully connected layer on the concatenation of their final embedding level. A simpler solution with just a sum of their normalized prediction scores provides close but lower quality results.

As another comparative method, we compare the results obtained with our approach with a baseline method relying on the bounding boxes of the objects of interest as in [27]. In this approach, the bounding box coordinates of each shape are encoded into 512-dimensional vectors by linear layers, and then are fused into one by element wise addition, which is then classified by a 2-layer fully connected network with 256 hidden units. In our solution (*bbox coords*), a Multi-Layer Perceptron (MLP) model is used directly on the coordinates of the two objects without considering them separately in the first stage, which allows more expressiveness. The network used is made of 4 fully connected layers, with 96, 192 and 96 hidden units, which results in 38 501 parameters with our 4 output classes.

3) Learning and validation protocol: Both CNN models were trained by fine-tuning a SqueezeNet model pre-trained on IMAGENET (Section III-C). The MLP used for the bounding box coordinates was trained from scratch. All the models were trained with cross-entropy loss and classic *SGD* optimizer, with a learning rate of 10^{-4} and default values for the other parameters ($\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 10^{-8}$), with a batch size of 20.

To avoid a learning bias and over-fitting, the datasets are classically split into two subsets with sizes of 75% and 25% representing respectively train and test sets, keeping the same proportions of each class and each ambiguity level as in the initial dataset in both subsets. Filtering of ambiguity levels can also be performed if only some levels have to be used.

Three trainings of each model were performed: the first one with all the *SimpleShapes* dataset (Section IV-A) without totally ambiguous cases (N1+N2+N3, 1953 images), the second one without the most difficult situations (N1+N2, 1794 images) and the last one with the simplest situations only (N1, 1421 images). So as to better evaluate the generalization capacity of each model, they were also tested on the full *GIS* dataset, which is made of different kinds of shapes. The training on simple cases of *SimpleShapes* (N1+N2 or N1 only) was also tested on difficult cases (N3) of the same dataset, as another evaluation of the generalization capacity, and of the difference of difficulty in practice between the different level subsets.

Table I
CLASSIFICATION RESULTS (OVERALL ACCURACY – OA AND STANDARD DEVIATION – STD) ON THE TEST SETS.

	dF-banner		*bbox image*		*dFB + image*		*bbox coords* [27]	
Datasets	**OA**	**STD**	**OA**	**STD**	**OA**	**STD**	**OA**	**STD**
Train & Test on N1	**92.66%**	0.94%	88.39%	0.50%	92.13%	0.25%	90.73%	1.71%
Train & Test on N1+N2	92.70%	0.62%	87.53%	0.46%	**92.90%**	0.78%	90.13%	0.34%
Train & Test on N1+N2+N3	91.47%	1.36%	87.30%	0.55%	**92.53%**	1.62%	88.96%	0.89%
Train on N1 & Test on N3	76.03%	3.76%	73.86%	1.15%	**78.13%**	2.65%	72.75%	3.46%
Train on N1+N2 & Test on N3	75.54%	2.70%	72.82%	1.54%	**78.55%**	2.79%	73.17%	0.96%
Train on N1+N2+N3 & Test on GIS	**91.81%**	0.79%	61.75%	3.65%	86.02%	2.43%	86.67%	3.17%

From each training, we selected the optimal number of epochs according to the loss obtained on the test part: the best models are those corresponding to the epochs when the loss starts to plateau or before it starts to increase. In our tests, we selected three following epochs and re-ran the test on the same test part three times for each one. Then we compute the mean over all the 3×3 runs, as an estimation of the accuracy that can be expected with those models, and the standard deviation to show the fluctuation of the performance for the different methods. The results are reported in Table I. This step can also be seen as a validation step for the test on the *GIS* dataset.

C. Results and discussion

As a preliminary result, Figure 5 shows a sample of discrete Force Banners for each class and for each level of ambiguity. For each class a privileged direction can be deduced, corresponding to the direction of the spatial relation. This is clear for the simple cases where the direction is close to the privileged one, while it gets fuzzy when it is too far from this privileged direction, or covering a range which is too large. These visual results first highlight the ability of the discrete Force Banner to capture (simple to more complex) spatial configurations in a directional manner.

Figure 6 illustrates the obtained loss, accuracy and F1-score curves (test/validation phase) when the models are trained with the different image representations on the *SimpleShapes* dataset. We observe that the different models start to provide their best results after 10 epochs, and that the test loss starts to plateau from 20 epochs for all models, and to increase from 30 for the *bbox image* model, or 50 for the *dF-banner* model, while it remains stable for the *bbox coords* model.

By looking at the accuracy or F1-score values from Figure 6, we can also compare the different approaches. For all the approaches, the scores start to converge after 10 epochs too, and get really stable after 45 epochs. The first one to reach its plateau is the *bbox coords* approach, which is already high after the first epoch (but has still important variation until 8 epochs), contrary to the others that start from $40 - 60\%$ and reach their plateau after 5 to 10 epochs. The best result is obtained for the *dF-banner* and *dFB+image* methods, followed by the *bbox coords* method just below, and finally the *bbox image* one which is quite inferior.

The precise results for this test and for the tests on the different datasets are reported in Table I, computed as mentionned in Section IV-B3. When training and testing on the same dataset, either the *dF-banner* or the *dFB+image* approach gets the best results, depending on the difficulty of the data, with a really small difference. They are followed by the *bbox coords* approach and finally the *bbox image* one in general.

Concerning the ambiguity level of the dataset, the difficulty of N3 relatively to N1 and N2 is clear from the tests on N3 only, with an important gap compared to tests on the same level of difficulty as in train. However the relative difference between training on N1 or on N1+N2 is not really visible on the results, which may be due to the higher importance of the N1 images in the dataset.

Finally for the test on the *GIS* dataset, the *dFB+image* score is much lower than on *SimpleShapes*, just below the *bbox coords* approach. This must be due to the bad performance of the *bbox image* approach, while the score for the *dF-banner* is still as high as for the test on the *SimpleShapes* dataset. This means that the *dF-banner* has a much higher and really good generalization capacity, since it can get higher performance on an unseen dataset corresponding to realistic situations.

All these results suggest that the discrete Force Banner is an appropriate description of the spatial configuration, providing better features for the classification task than the bounding box coordinates or the binary image, even if the latest can bring some additional information in the more difficult cases. It is particularly good at generalizing to other kind of images, as the test on the *GIS* dataset reveals.

V. CONCLUSION

In this article, we studied the problem of spatial relation recognition. Instead of considering directly the original image space to predict the spatial relation, we proposed the Force Banner representation modeling rich spatial information between pairs of objects composing a scene. Such an intermediate representation captures the relative position between objects using a panel of forces (attraction and repulsion), that take into account the structural shapes of the objects and their distance in a directional fashion. This solution currently only deals with binary images, but it is robust to imperfect segmentation since the contribution of one pixel is balanced by the others for each direction. Moreover, Force Banners can be used to feed a classical $2D$ CNN, here the SqueezeNet model, for the recognition of spatial relations, benefiting from pre-trained models and fine-tuning.

Experimental results obtained on a dataset of about 2000 images composed of various shapes (simple objects, geomet-

rical shapes, urban objects from a GIS) highlighted the interest of this approach, and in particular its benefit to describe spatial information, with very good generalization capacity. We also shown in this study that current state-of-the-art approaches based on bounding-boxes models may not be sufficient to predict spatial relations from ambiguous or complex relations. Thus, we provide here research clues to allow a CNN to better apprehend the spatial information carried naturally by the content of a scene.

As a perspective, we plan to embed the Force Banners in a graph-based representation of the image to allow the recognition of complex spatial configurations between scenes composed of multiple objects. We also want to integrate in our approach a post-hoc visual attention mechanism, to better understand and visualize which panel of forces (repulsions, attractions) and which set of directions contributed to the decision of the CNN classifier. This will ultimately allow to refine and adapt the model to the specificity and complexity of the classes from a dataset.

ACKNOWLEDGEMENTS

This work was carried out at the LIPADE and funded by Magellium, with the support of the French Defense Innovation Agency (AID).

REFERENCES

[1] P. Matsakis and L. Wendling, "A New Way to Represent the Relative Position between Areal Objects," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 7, pp. 634–643, 1999.

[2] I. Bloch, "Fuzzy spatial relationships for image processing and interpretation: A review," *Image and Vision Computing*, vol. 23, no. 2, pp. 89–110, 2005.

[3] P. Matsakis, L. Wendling, and J. Ni, "A General Approach to the Fuzzy Modeling of Spatial Relationships," in *Methods for Handling Imperfect Spatial Information*, 2010, pp. 49–74.

[4] I. Bloch, "Fuzzy Relative Position between Objects in Image Processing: A Morphological Approach," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 7, pp. 657–664, 1999.

[5] R. M. Cesar, E. Bengoetxea, and I. Bloch, "Inexact graph matching using stochastic optimization techniques for facial feature recognition," in *Int. Conf. on Pattern Recognition (ICPR)*, vol. 2, 2002, pp. 465–468.

[6] O. Colliot, O. Camara, and I. Bloch, "Integration of fuzzy spatial relations in deformable models - Application to brain MRI segmentation," *Pattern Recognition*, vol. 39, no. 8, pp. 1401–1414, 2006.

[7] A. Delaye and E. Anquetil, "Learning of fuzzy spatial relations between handwritten patterns," *International Journal on Data Mining, Modelling and Management*, vol. 6, no. 2, pp. 127–147, 2014.

[8] K. Miyajima and A. Ralescu, "Spatial organization in 2D segmented images: Representation and recognition of primitive spatial relations," *Fuzzy Sets and Systems*, vol. 65, no. 2, pp. 225–236, 1994.

[9] X. Wang and J. M. Keller, "Human-based spatial relationship generalization through neural/fuzzy approaches," *Fuzzy Sets and Systems*, vol. 101, no. 1, pp. 5–20, 1999.

[10] P. Matsakis, J. M. Keller, L. Wendling, J. Marjamaa, and O. Sjahputera, "Linguistic description of relative positions in images," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 31, no. 4, pp. 573–88, 2001.

[11] A. R. Buck, J. M. Keller, and M. Skubic, "A memetic algorithm for matching spatial configurations with the histograms of forces," *IEEE Transactions on Evolutionary Computation*, vol. 17, no. 4, pp. 588–604, 2013.

[12] S. Tabbone and L. Wendling, "Color and grey level object retrieval using a 3D representation of force histogram," *Image and Vision Computing*, vol. 21, no. 6, pp. 483–495, 2003.

[13] M. Clément, M. Garnier, C. Kurtz, and L. Wendling, "Color object recognition based on spatial relations between image layers," in *Int. Conf. on Computer Vision Theory and Applications (VISAPP)*, 2015, pp. 427–434.

[14] M. C. Vanegas, I. Bloch, and J. Inglada, "A fuzzy definition of the spatial relation "surround" - Application to complex shapes," in *European Society for Fuzzy Logic and Technology (EUSFLAT)*, 2011, pp. 844–851.

[15] I. Bloch, O. Colliot, and R. M. Cesar, "On the Ternary Spatial Relation "Between"," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 36, no. 2, pp. 312–327, 2006.

[16] N. Loménie and D. Racoceanu, "Point set morphological filtering and semantic spatial configuration modeling: Application to microscopic image and bio-structure analysis," *Pattern Recognition*, vol. 45, no. 8, pp. 2894–2911, 2012.

[17] M. Clément, A. Poulenard, C. Kurtz, and L. Wendling, "Directional Enlacement Histograms for the Description of Complex Spatial Configurations between Objects," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2366–2380, 2017.

[18] P. Matsakis, M. Naeem, and F. Rahbarnia, "Introducing the Φ-Descriptor – A Most Versatile Relative Position Descriptor," in *Int. Conf. on Pattern Recognition Applications and Methods (ICPRAM)*, 2015, pp. 87–98.

[19] P. Matsakis and M. Naeem, "Fuzzy Models of Topological Relationships Based on the PHI-Descriptor," in *IEEE Int. Conf. on Fuzzy Systems (FUZZ-IEEE)*, 2016, pp. 1096–1104.

[20] J. F. Allen, "Maintaining knowledge about temporal intervals," *Communications of the ACM*, vol. 26, no. 11, pp. 832–843, 1983.

[21] M. Clément, C. Kurtz, and L. Wendling, "Bags of spatial relations and shapes features for structural object description," in *IEEE Int. Conf. on Pattern Recognition (ICPR)*, 2016, pp. 1995–2000.

[22] M. Clément, C. Kurtz, and L. Wendling, "Learning spatial relations and shapes for structural object description and scene recognition," *Pattern Recognition*, vol. 84, pp. 197–210, 2018.

[23] J. Peyre, I. Laptev, C. Schmid, and J. Sivic, "Weakly-supervised learning of visual relations," in *IEEE Int. Conf. on Computer Vision (ICCV)*, 2019, pp. 5189–5198.

[24] B. Dai, Y. Zhang, and D. Lin, "Detecting visual relationships with deep relational networks," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 3298–3308.

[25] H. Zhang, Z. Kyaw, S. Chang, and T. Chua, "Visual translation embedding network for visual relation detection," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 3107–3115.

[26] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L. Li, D. A. Shamma, M. S. Bernstein, and L. Fei-Fei, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *International Journal of Computer Vision*, vol. 123, no. 1, pp. 32–73, 2017.

[27] K. Yang, O. Russakovsky, and J. Deng, "Spatialsense: An adversarially crowdsourced benchmark for spatial relation recognition," in *IEEE Int. Conf. on Computer Vision (ICCV)*, 2019, pp. 2051–2060.

[28] J. Freeman, "The Modelling of Spatial Relations," *Computer Graphics and Image Processing*, vol. 4, no. 2, pp. 156–171, 1975.

[29] A. Rosenfeld and R. Klette, "Degree of adjacency or surroundedness," *Pattern Recognition*, vol. 18, no. 2, pp. 169–177, 1985.

[30] F. Iandola, M. Moskewicz, K. Ashraf, S. Han, W. Dally, and K. Keutzer, "Squeezenet: AlexNet-level accuracy with 50x fewer parameters and <1MB model size," *Computing Research Repository*, vol. abs/1602.07360, 2016.