Representation and Retrieval of Images by Means of Spatial Relations Between Objects

Danilo Nunes
Leonardo Anjoletto Ferreira
Paulo Eduardo Santos
Centro Universitário da FEI
São Bernardo do Campo, Brazil
nunesdanilo@gmail.com,{psantos, laferreira}@fei.edu.br

Adam Pease
Infosys,
Foothill Research Center
Palo Alto, CA, USA
adam.pease@infosys.com

Abstract

The present work addresses the challenge of integrating low-level information with high-level knowledge (known as semantic gap) that exists in content-based image retrieval by introducing an approach to describe images by means of spatial relations. The proposed approach is called Image Retrieval using Region Analysis (IRRA) and relies on decomposing images into pairs of objects. This method generates a representation composed of n triples, each one containing: a *noun*, a *preposition* and, another *noun*. This representation paves the way to enable image retrieval based on spatial relations. Results for an indoor/outdoor classifier shows that neural networks alone are capable of achieving 88% in precision and recall, but when combined with ontology this result increases in 10 percentage points, reaching 98% of precision and recall.

1 Introduction

In this work we investigate the application of spatial relations in the image retrieval problem. The issue of representing the semantics existent in an image has been receiving great attention recently. Numerical methods (low level) are not able to fully integrate semantics (high level) due to the fact that the semantic content might be constituted by qualitative concepts. The challenge of integrating low-level information with high-level knowledge is a known problem in computer vision, often referred as the semantic gap. In this work, a multi-level approach, called *Image Retrieval using Region Analysis (IRRA)*, is proposed to retrieve images by their semantics from a bottom-up knowledge representation procedure. The method proposed here ensembles a stack of distinct neural networks in order to estimate spatial relations, expressed by a spatial preposition between pairs of objects. This procedure permits a representation of an image by the objects depicted and the relations holding between them. Thus, a sparse representation is constructed taking into account these objects and their relations. The images are indexed based on this sparse representation in order to enable fast retrieval. Finally, we extend a public ontology with this data in order to infer new relations beyond the original binary relations. This representation enables the retrieval of images based on queries with respect to spatial arrangements.

To semantically interpret an image it is necessary to: define the context, detect objects, define some similarity metric and, finally, apply some method for knowledge representation (Wan et al. 2014). We can organize the proposed framework in two distinct steps: the quantitative analysis, related to low-level information processing, and the qualitative analysis, which stands for high-level knowledge representation.

The quantitative analysis tackles the semantic gap problem using a hierarchy of classifiers. We represent the semantics building a top-down approach which contains a specific classifier for each of the following tasks: scene recognition, object segmentation and preposition estimation. The proposed approach decomposes an image into scenes and then, for each scene, it segments the related objects; given a pair of these objects, the method estimates a spatial preposition.

The qualitative analysis is built on top of Suggested Upper Merged Ontology (SUMO)[1] (Niles and Pease 2001; Pease 2011) and is constructed with the data obtained at the quantitative phase, i.e., scenes, objects and prepositions. This representation contains the distinct segmented objects and their relations, expressed by a spatial preposition.

For evaluation purposes, the method proposed is applied to image retrieval tasks. The results obtained show that our method outperforms recent approaches aiming at the retrieval of images by means of spatial relations. Results also show that the classification task is much improved with a combination of neural networks (working at low-level information) with an object ontology (representing high-level knowledge) in contrast to using the neural network classifiers alone.

2 Related work

One of the most common approaches for retrieving images is the paradigm known as query by example. Bag of Visual Words (BOVW) is a method widely used to perform image retrieval in this context. BOVW extracts local features and, with respect to a sparse representation, performs image retrieval (Philbin, Sivic, and Zisserman 2008). BOVW-

[1] www.ontologyportal.org

based techniques retrieve images by their visual similarity. However, the retrieval task does not take into account the meaning (or semantics) of the information sought, since it is solely based on numerical analyses. On the one hand, the way an image is represented is crucial to enable a fast retrieval of images. On the other hand, there is poor correlation between the semantics one image might contain and its constructed representation (Hudelot, Atif, and Bloch 2008).

One possible way to represent the semantics expressed in images is by following the steps described in (Hare et al. 2006) extract and describe the features of interest, perform object segmentation and create a high-level representation of the detected regions in the images. Considering non-structured regions in images, the goal is to assign a distinct label to each one of them. A number of approaches have focused in labeling regions in an image. To name some of the most relevant we can refer to those based on Conditional Random Fields (CRF) (Gould, Fulton, and Koller 2009), approaches based on deep Neural Networks (NN) (Socher et al. 2011) or Convolutional Neural Networks (CNN) (Girshick et al. 2014), combinations of CRF and CNN have also been applied to this task (Zheng et al. 2015). These methods, however, do not take into account relations between objects in images, as considered in this paper. It is worth noticing that, in order to reduce the semantic gap, it is important to represent the spatial relations among objects, since these relations suffer less from viewpoint changes than the object recognition itself (Bloch, Hudelot, and Atif 2007). Some work has investigated the inclusion of high-level knowledge focused on spatial relations in image analysis. For example, we can cite the work of (Bloch, Hudelot, and Atif 2007) and (Hudelot, Atif, and Bloch 2008). Both of them establish high-level assumptions to enhance the low-level processing. And more recently, we refer to (Lu et al. 2016) and (Dai, Zhang, and Lin 2017) both of them establish relations between objects in an image based on a set of distinct predicates. Similarly, (Malinowski and Fritz 2014) focus on image retrieval by using queries built on spatial relations between objects in images and (Mai et al. 2017) that retrieves images based on spatial arrangement of an example input image. In the present paper, we propose a new paradigm that combines quantitative processing and also qualitative analysis in an end-to-end architecture.

In order to establish a spatial relation, it is necessary to define a reference system. For instance, considering the relation x is ***in front of*** y, three concepts should be defined: the target object, the reference object and the reference system (Hudelot, Atif, and Bloch 2008). The reference system is, in general, categorized from the observer's viewpoint (relative or absolute), or with respect to the way that the relation is used: intrinsic, extrinsic or deictic. This work uses intrinsic relations, which specify a relation under the perspective of an observer. It is important to reinforce that these relations are not constant in time and can also change their status with respect to the adopted perspective. The present work uses spatial relations as defined by a general ontology in order to accomplish the automatic extraction of the semantic content present in an image.

There are distinct definitions of ontology found in the literature (Sankat, Thakur, and Jaloree 2016). In this work, ontology is understood as in (Lehmann and Vöelker 2014) which defines ontology as a formal knowledge representation, which may or may not be restricted to a specific domain. This representation is expressed in a manner that might be understood by a computational process (Sankat, Thakur, and Jaloree 2016). Ontology might be referred as the commonsense knowledge with respect to a domain of interest, and can be expressed by: concepts, relations between concepts, functions and instances. The Suggested Upper Merged Ontology (SUMO), used in this work, extends these concepts with axioms in higher-order logic that attempt to define each concept.

3 Image Retrieval using Region Analysis (IRRA)

In this work we address the problem of the semantic gap in image retrieval by representing an image by the spatial relation among the objects existing in the image itself. Our aim is to combine information obtained from the pixel level (quantitative analysis) with information provided by specialists (qualitative analysis), in this work the specialist knowledge is defined in terms of spatial prepositions in a general ontology. The method proposed in this work is called Image Retrieval using Region Analysis (IRRA) and is summarized in Figure 1. Considering an input image, IRRA applies a neural network in order to segment objects in the image, that are further combined into pairs of objects. In the sequence, for each pair, a preposition between them is estimated. Thus, we decompose the image into n-triples containing pairs of objects and a spatial preposition. These n-triples could, thus, be used to represent (and reason about) the domain. Section 3.1 describes the quantitative analysis of IRRA in more detail.

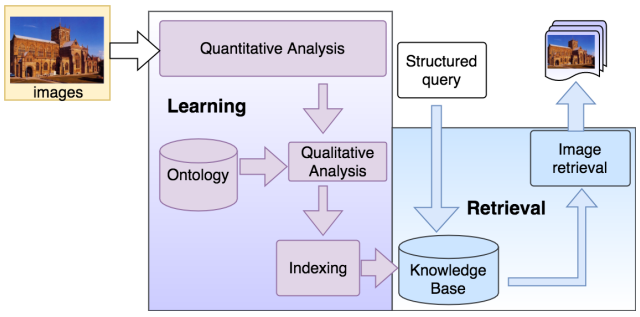


Figure 1: Overview of IRRA method numerical analysis phase.

3.1 Quantitative analysis

The proposed pipeline of the quantitative analysis is the following:

1. Identify the context in which objects are inserted;
2. Detect regions in the images occupied by each object;
3. Combine the detected objects in pairs;

4. Estimate a spatial relation for every detected object pair.

In order to perform all these tasks, a stack of neural networks is applied. We apply three distinct neural networks to hierarchically detect context, objects and spatial relations.

Context identification: The first neural network is applied in order to estimate context. In this work, context is understood as the scene class an image belongs to. Two scene classes are used in this work: *indoor* and *outdoor*. We created a model whose architecture is identical to that described in AlexNet (Krizhevsky, Sutskever, and Hinton 2012) in order to perform this classification. The purpose of this step is to reduce the number of target objects, dismissing inconsistent objects with respect to the scene; objects such as buildings and cars will not be part of the segmentation of an indoor model, for instance.

Object segmentation: Object (semantic) segmentation is the second step to represent high-level information in an image. This task is challenging since we want to classify an image pixel-wise; besides, object identification is dependent on the context that the objects are immersed in. The neural networks that are performing semantic segmentation in this work have the same architecture, although each one has its own set of weights. The main difference between them is the target classes that they use to construct the model.

The segmentation neural networks generate a set of class proposals for each pixel in the image. Additionally, similar pixels are grouped together in order to represent objects. In this work we focus on assigning spatial relations between pairs of these segmented objects. In order to define a spatial region, we map pairs of detected objects aiming at establishing a representation that might be expressed in topological terms. Then, another neural network is used to estimate a spatial preposition from the previous (topological) classification. Each image I_x generates a $C^{I_x} = \binom{n^{I_x}}{2}$, where n^{I_x} is the number of objects detected for image I_x, and C^{I_x} is the combination of n^{I_x} in pairs.

Preposition estimation: A sparse representation of C^{I_x} is created and is provided as input to a neural network whose task is to assign English prepositions to images. It is important to notice that object information is crucial at this point to disambiguate multiple (possible) assignments. Besides, this sparse representation provides a notion of the spatial topology with respect to a pair of objects.

Finally, this sparse vector combined with the estimated spatial preposition enables the creation of an index that can be accessed with a structured query in order to retrieve relevant images. This index is similar to those used by BOVW(Csurka et al. 2004) applications. Thus, images can be retrieved given a feature which, in the present case, is a spatial relation.

3.2 Qualitative analysis

Knowledge representation is built with respect to the spatial arrangement of the identified structures in images. The interpretation of spatial relations contributes to reducing the semantic gap in images since relations tend to suffer less with variations than the objects in their arguments (Hudelot, Atif, and Bloch 2008).

Aiming at representing the knowledge acquired through the numerical analysis of images, we have opted to extend an existing ontology: SUMO (Pease 2011). The Suggested Upper Merged Ontology (SUMO) has been adopted due to the fact it is built on higher-order logic and, thus, it provides the flexibility to create complex constructions, not restricted to binary relations.

The initial step to represent knowledge was to create instances of the domain. In this work, each detected object is considered as an unique and independent instance.

Considering an example image I_x with the detected objects *Building, Floor* and *Sky*, in this work, each of these terms is identified by the suffix corresponding to the original image, in this case x. Therefore the term *Building* is referenced by $Building_x$ and so on. Formula 1 shows the instance definition.

$$
\begin{array}{l}
(\text{instance}\ \ Building_x\ \ \text{Building}) \\
(\text{instance}\ \ Floor_x\ \ \text{Floor}) \\
(\text{instance}\ \ Sky_x\ \ \text{Sky})
\end{array}
$$

Formula 1: Instances

All the images and the detected objects are represented in a similar fashion. The next step of our framework is to construct spatial relations between the distinct objects. To execute the mapping of these relations to the segmented pair of objects, we have considered the prepositions estimated by a statistical classifier. Finally, our preposition domain contains the following relations: *above, across from, behind, below, in, in front of, inside of, left of, on, right of* and *under*.

By using an off-the-shelf ontology, the representation of these relations expressed by prepositions was simplified. In this work, only a punctual extension of SUMO was necessary.

The standard SUMO ontology provides tools to define spatial relations. In order to define a spatial relation in SUMO it is necessary to create an instance of the class *PositionalAttribute*. This class enables the stating of binary orientation relations between two objects. Additionally, it is important to mention that the semantics with respect to spatial relations might be expressed by more than one preposition, besides, distinct prepositions might be similar or complementary. For instance, two distinct objects arranged consecutively might be referenced by the following prepositions : *in front of* or *behind*, the proper term is defined according to the context and the observer's position. This characteristic allows us to represent both relations using the double implication operator $<=>$. Consequently, it is possible to define the preposition *in front of* based on the preposition *behind*, or vice-versa. This definition is illustrated in Formula 2. It is worth pointing out that, although *in front of* or *behind* are both relative to an observer, this paper assumes that observer is the camera point of view and, thus, it is implicit in the definitions.

$$
\begin{array}{l}
(<=> \\
\quad (\text{orientation}\ \ ?X1\ \ ?X2\ \ \text{Behind}) \\
\quad (\text{orientation}\ \ ?X2\ \ ?X1\ \ \text{InFrontOf}))
\end{array}
$$

Formula 2: Double implication to *behind* and *In front of*

Table 1: Precision-recall for scene classification.

Scene	Precision	Recall	N
Indoor	0.86	0.84	1,829
Outdoor	0.88	0.90	2,439
Overall	0.87	0.87	4,268

According to the nature of spatial prepositions, this procedure is applied to other relations, for instance: *under* and *above* or *left of* and *right of*. Additionally, the transitivity of relations was also used in this context whenever possible. For instance, we might infer that if there is an object a above an object b and b is above a third object c, therefore a is above c.

In conclusion, through quantitative methods, qualitative information with respect to the domain was inferred. The impact of this procedure was evaluated in the tests described in the next section.

4 Experiments

This section details the experiments executed in order to evaluate the proposed method. To perform the experiments, the publicly available data set SUN09 (Choi et al. 2010) was used. This data set is composed of more than 12,000 images containing various classes of objects in distinct scenes. In order to conduct the tests, two data sets of annotations for SUN09 were used. First, the data set provided by (Lan et al. 2012) (data set 1) was used, that contains annotations in the form of structured queries ($\langle noun, preposition, noun \rangle$) representing two relations $below$ and $above$. Second, the annotations provided in (Malinowski and Fritz 2014) (data set 2) were considered that includes eleven (11) distinct preposition classes: *above, across from, behind, below, in, in front of, inside, left, right, on* and *under*.

Overall 4,367 images were used for training and 4,317 images for testing. In these datasets there are 186,299 pairs of objects for training and 173,111 for testing, being 106 distinct objects considering data set 1 and 42 considering data set 2.

Table 2: Intersection over union.

Scene	Intersection Over Union %	
	Context	No context
Outdoor	24.11	18.60
Indoor	16.57	14.07
Average	**22.66**	**18.67**

4.1 Scene classification

The scene classification part of the present proposal (identifying what we use as context: *indoor* or *outdoor* scenes) was evaluated using data set 2, with manual annotations for training and evaluation purposes. The results of this binary classification are shown in Table 1, where we can see that the overall precision and recall for each of the considered classes was 87%, attesting for the suitability of the method applied for this task.

As described above, scene identification provides information to refine the object segmentation procedure, whose results are shown below.

4.2 Object segmentation

In this part of the system evaluation, we investigate the hypothesis of whether the information provided by the scene identification (Section 4.1) improves the segmentation. According to this premise we have manually separated the objects as indoor and outdoor. The indoor objects are: *armchair, basket, bookcase, book, bottle, box, chair, closet, cupboard, curtain, desk, floor, flower, ground, mirror, plant, poster, refrigerator, seats, table, vase, wall and window*. The outdoor objects are: *airplane, balcony, bench, building, car, door, fence, gate, grass, path, road, rock, sign, sky, streetlight, tree, van, water*. According to each scene class detected we apply one or the other segmentation models (i.e. one trained with *indoor* objects or the other trained with *outdoor objects*).

The semantic segmentation model applied in this experiment was fine tunned with the weights provided by (Zhou et al. 2016). The object segmentation results are shown in Table 2. Table 2 shows all the objects in data set 2. In the leftmost column is the scene class of which the object belongs to. The second column, from left to right, is the noun that represents the object. The following two columns represent the Intersection over union (IoU) (Jaccard 1912) considering object recognition using context information or without using it (column "No context"). Considering the labeled area A_L for a given object in an scene and the segmented area A_S resulting from a neural network segmentation, the Intersection over Union is the ration between the intersection of A_S and A_L and the union between the regions, as shown in equation 1.

$$\text{IoU} = \frac{A_S \cap A_L}{A_S \cup A_L} \tag{1}$$

thus, when the neural network segments exactly the same area as the label ($A_S = A_L$), $\text{IoU} = 1$ and when the segmented area A_S is completely different from the label area A_L ($A_S \cap A_S = \emptyset$), then $\text{IoU} = 0$. For any situation when the segmented area overlaps with the labelled area ($A_S \cap A_S \neq \emptyset$) but are not equal, $0 < \text{IoU} < 1$.

Results show that the classification using context information had a higher IoU value overall (shown in line "Average" in Table 2). However, there were various cases in which the use of scene class information did not improve the results (such as the results related to *Armchair*, or *Bench*), this occurred due to the fact that such objects appear in both (indoor and outdoor) scenes. Thus, in these cases, assuming scene classes caused a reduction in the number of examples of some objects during the training phase.

4.3 Scene classification with segmentation and ontologies

With the object-segmentation neural network presented in the previous section, a second *indoor/outdoor* classifier was

tested, in which we combine low-level information of the scene with high-level information provided by the description of the objects in SUMO.

For this classifier, each of the objects presented in Section 4.2 was described in SUMO as having one of three possible classes: *indoor* (e.g., *armchair*), *outdoor* (e.g., *sky*) or *both* (for objects that are both *indoor* and *outdoor*, e.g., *chair*). The complete list of object classes and the corresponding SUMO annotations are shown in Table 3. With the same segmentation method as used in Section 4.2, the objects recognised in each of the scenes were counted and the class with the highest amount of objects present was considered to be the scene class (e.g., a scene with 2 *indoor* objects, 1 *outdoor* and 1 *both* is considered to be an *indoor* scene). Scenes where there is an equal number of *indoor* and *outdoor* objects were classified as *both*. In this case, the object rank was used in order to take into account the relevance of the object with respect to the scene classification. For instance, a scene containing an object "sky" is considered to be and outdoor scene, since there can be no sky indoors[2]). We considered that the following *outdoor* objects outrank any other object in the domain: *building*, *sky*, *sign*, *fence*, *grass*, *road*.

Results obtained for this combination of low-level and high-level classification are shown in Table 4 where we obtained over 98% of precision and recall. An improvement of 10 percentage points with respect to the classifier without the ontology (whose results are shown in Table 1).

From 4,268 scenes used in this experiment 44 could not be classified into any of the three classes available, since they did not contain any object, or contained only one object recognised in the class *both*. These were not accounted in Table 4.

4.4 Preposition assignment

The main difficulty of assigning a preposition to a pair of objects in an image is the common overlapping of terms, i.e., there are several possible (consistent) preposition assignments to each spatial relation. In order to cope with this issue, before assigning a preposition, the topological relation between pairs of objects is classified, serving as a bridge to preposition definition. In our experiments, each image in the data set was transformed into a representation containing: *(target object, reference object)*. This representation is shown in Figure 2, Figure 2a presents the original image and the object mask is presented in Figure 2b. It is important to mention the fact that the color of the objects indicates the target object (blue) and the reference object (red).

Each image available in the data set was segmented and combined with its relative object, generating a combination of $C_{I^x} = \binom{N_{I^x}}{2}$ for images I^x and segmented objects N_{I^x} for the image. Every image generated from C_{I^x} was classified according to the spatial preposition related to the reference and target objects. In this test we have used data set 2 with 11 spatial prepositions.

[2]In this experiment, a window is recognised as a single object, therefore no other object can be perceived within its contours.

Table 3: Objects' classes and SUMO annotation

Object	Class	SUMO Annotation
Background	Both	(subclass Background Both)
Airplane	Outdoor	(subclass Airplane Outdoor)
Armchair	Indoor	(subclass Armchair Indoor)
Balcony	Outdoor	(subclass Balcony Outdoor)
Basket	Indoor	(subclass Basket Indoor)
Bench	Outdoor	(subclass Bench Outdoor)
Bookcase	Indoor	(subclass Bookcase Indoor)
Books	Indoor	(subclass Books Indoor)
Bottle	Indoor	(subclass Bottle Indoor)
Box	Indoor	(subclass Box Indoor)
Building	Outdoor	(subclass Building Outdoor)
Car	Outdoor	(subclass Car Outdoor)
Chair	Indoor	(subclass Chair Indoor)
Closet	Indoor	(subclass Closet Indoor)
Cupboard	Indoor	(subclass Cupboard Indoor)
Curtain	Indoor	(subclass Curtain Indoor)
Desk	Indoor	(subclass Desk Indoor)
Door	Both	(subclass Door Both)
Fence	Outdoor	(subclass Fence Outdoor)
Floor	Indoor	(subclass Floor Indoor)
Flower	Both	(subclass Flower Both)
Gate	Both	(subclass Gate Both)
Grass	Outdoor	(subclass Grass Outdoor)
Ground	Outdoor	(subclass Ground Outdoor)
Mirror	Indoor	(subclass Mirror Indoor)
Path	Outdoor	(subclass Path Outdoor)
Plant	Both	(subclass Plant Both)
Poster	Indoor	(subclass Poster Indoor)
Refrigerator	Indoor	(subclass Refrigerator Indoor)
Road	Outdoor	(subclass Road Outdoor)
Rock	Outdoor	(subclass Rock Outdoor)
Seats	Indoor	(subclass Seats Indoor)
Sign	Outdoor	(subclass Sign Outdoor)
Sky	Outdoor	(subclass Sky Outdoor)
Streetlight	Outdoor	(subclass Streetlight Outdoor)
Table	Indoor	(subclass Table Indoor)
Tree	Outdoor	(subclass Tree Outdoor)
Van	Outdoor	(subclass Van Outdoor)
Vase	Indoor	(subclass Vase Indoor)
Wall	Indoor	(subclass Wall Indoor)
Water	Both	(subclass Water Both)
Window	Both	(subclass Window Both)

Table 5 shows the precision, recall and f-measure for each of the tested prepositions, where the right-most column is the number of tested samples. The overall values are exhibited in the last line: for the 4,953 relations tested, the system reached an overall precision of 0.75 with a recall of 0.75 and a f-measure of 0.71.

Next section compares the performance of IRRA with other state-of-the-art methods in the task of image retrieval from structured queries.

4.5 Retrieval evaluation

In order to evaluate image retrieval using IRRA, the annotations provided by (Lan et al. 2012) were used. To the best of our knowledge, this is the only available data set that maps spatial relations to objects detected in images. We have

Table 4: Precision-recall for classification with ontology.

Scene	Precision	Recall	N
Indoor	0.98	0.98	1,813
Outdoor	0.99	0.98	2,463
Overall	0.99	0.98	4,320

Table 5: Precision, recall and f-measure for the estimated prepositions.

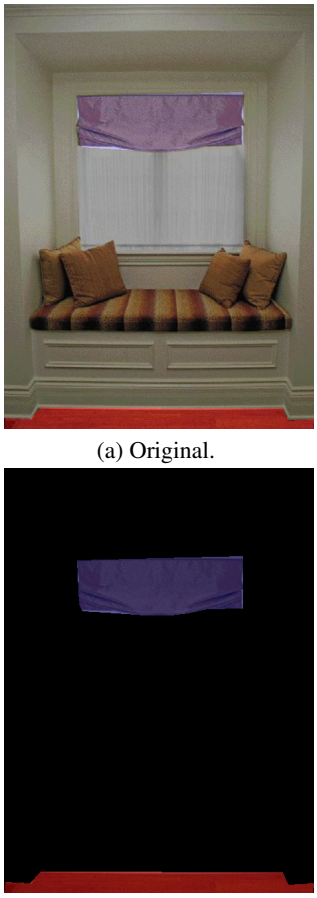
Preposition	precision	recall	f-measure	n
Above	0.76	0.58	0.66	166
Across from	1.00	0.03	0.06	387
Behind	0.65	0.62	0.63	329
Below	0.84	0.79	0.81	361
In	0.59	0.84	0.69	475
In front of	0.55	0.69	0.61	317
Inside of	0.00	0.00	0.00	65
Left of	0.01	0.01	0.01	187
On	0.60	0.77	0.67	208
Right of	0.00	0.00	0.00	59
Under	0.87	0.98	0.93	2,399
Overall	0.75	0.75	0.71	4,953

tested our method against all the structured query types proposed in (Lan et al. 2012). The structured queries contain a *noun*, e.g. *pedestrians*, or a *relation* set expressed by a triple in the form *(noun, preposition, and noun)*, e. g. *"car on the road"*. The available structures are represented as: *Structure a (Sa)*, which contains only a relation set, for instance, *"car on road"*; *Structure b (Sb)* contains a relation set and a noun, e.g, *"car on road, pedestrians"*; *Structure c (Sc)* contains two relation sets, e.g., *"car on road, sky above building"*; *Structure d (Sd)* contains two relation sets and a noun, e.g., *"car on road, sky above building, pedestrians"*; *Structure e (Se)* contains three relation sets, e.g., *"car on road, sky above building, books inside of bookcase"*.

Figure 3 illustrates the obtained results and also displays a comparison with other approaches using the same data set. Only recall is presented since this is the measurement used in (Lan et al. 2012). According to Figure 3 it is possible to observe that IRRA outperforms the other methods in all the structured queries considered.

Analyzing IRRA results alone, we can see that in larger queries that do not include single objects (in scenarios *Sc* and *Se*, for instance), IRRA's performance is not as good as in other scenarios. This behavior occurs due the fact that, when answering a query such as *Sc* or *Se*, segmentation or preposition detection errors are propagated to the retrieval task.

We have also evaluated the retrieval by using the second data set which has a larger set of prepositions. In this setting IRRA achieved a retrieval with mean average precision (mAP) of 53.95, outperforming the recent results reported in (Malinowski and Fritz 2014). The superior performance of IRRA with respect to other (competing) methods is explained by the fact that IRRA uses the various relations existing in an image in the retrieval task.



(a) Original.

(b) Object mask.

Figure 2: Topology representation with highlighted objects.

4.6 Ontology expansion

The queries executed in the tests above were strictly unidirectional due to the fact that there is no high-level reasoning with respect to the spatial arrangement of the items in the document collection. To address this issue, we investigated the application of reasoning using SUMO.

We extended the annotations in (Malinowski and Fritz 2014) with the aim to infer new relations derived from the original relations that were manually annotated. Through the use of the SUMO ontology we extended the system's knowledge about the relations in order to evaluate spatial prepositions that were not obtained by the quantitative analysis processes.

In order to evaluate this method, distinct queries were proposed, but keeping the same information to be retrieved from the set of images. To accomplish this task inverse relations were applied. For instance, for two objects (x, y) and the query (x-above-y), we also evaluated the retrieval for (y-below-x) against the same annotation for above as used in the original query.

The instanciation of the whole ontology with SUN09 data

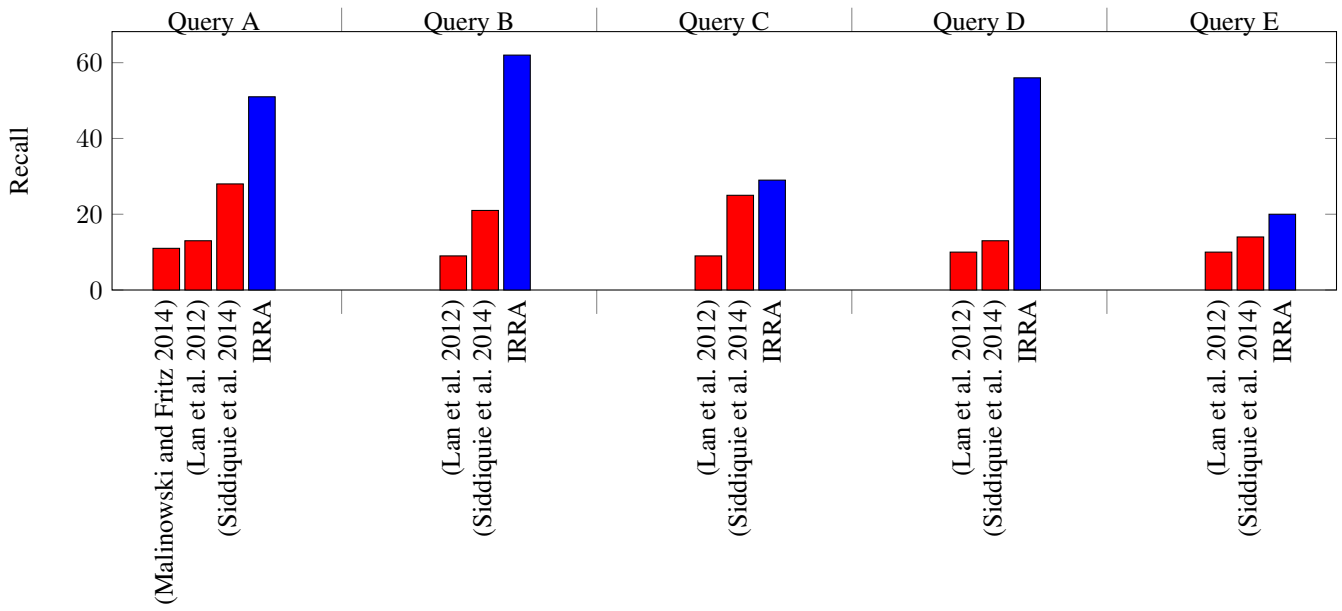


Figure 3: Comparison with other approaches.

generated more than $13,000$ terms with respect to the images and objects, and more than $18,000$ formulas referencing the created relations. The retrieval task based on the new set of queries was performed by evaluating every image using the E first-order logic theorem prover (Schulz 2013). Although SUMO is defined in a higher-order logic, we were able to achieve our goals with just the first-order logic content of the theory in this work, which allowed us to use a first-order logic prover.

Every image was tested using the new annotated queries, the mAp achieved in this case was 41.60, whereas he original set for these prepositions obtained a mAp of 51. The decrease in performance observed with the extended set of relations (in contrast with the original) was due to the fact that, by increasing the size of the knowledge base, errors were possibly included in the process, whose detection becomes increasingly complex to perform (Pease 2011).

5 Discussion

This paper proposed a framework for image retrieval based on an ensemble of neural networks and spatial relations defined over an ontology. The proposed method decomposes the image in distinct levels in order to classify objects and their spatial relations in static scenes. The retrieval of an image is than done by specifying the existence of objects and spatial relations between two objects (e.g., "car on road, pedestrians" is a possible sentence to use to retrieve an image).

Although the number of objects and relations seems to be small (42 objects and 11 relations), the number of images used is over 9,000 and since a scene is divided in relations between objects in it, it can be used for training more than one relation and/or terms, thus over 186,000 pairs of objects in training and 173,000 for testing were generated. This resulted in more than 13,000 terms and 18,000 formulas for the ontology.

Results on preposition classification also show that the method proposed in this paper outperforms previous work in the retrieval of images using spatial relations, however, it did not perform as well as expected in the following cases: *across from, inside of, left of, right of*. This problem was probably due to the distinct competing prepositions that could be equally applied to the scenes in these cases. A deeper investigation of how to represent the object pairs in order to enhance the estimation of their spatial preposition is a task for future work. It is also within our future interests to augment the set of relations covered by the system, relaxing the present restriction to binary relations.

When classifying an image as an *indoor* or *outdoor* scene using ontology, it is possible to specify an objects class, e.g., (subclass Sky Outdoor) or describe it in relation to another object, e.g., (subclass Armchair Chair), making it easier to add new objects to the classifier and to reason about their properties than it is to train a new deep learning method to classify new objects in the database. As a result, an increase of 10 percentage points was observed in the results when using the ontology to classify object's class found in the outputs of the neural networks, suggesting a successful combination of a knowledge representation tool with a state-of-the-art machine learning algorithm with virtually no learning or classification running time increase.

6 Conclusion

This work investigated the semantic gap that exists in content-based image retrieval by introducing an approach that establishes relations between objects in images by means of spatial arrangements. The method proposed in this paper, called Image Retrieval using Region Analysis

(IRRA), starts by decomposing images with respect to pairs of objects, where each pair is also combined with a spatial relation. Each spatial relation is related to a spatial preposition expressed in natural language. IRRA was evaluated on a public data set, whose results show that our approach outperforms previous (recent) work in the retrieval of images using spatial relations.

Results showed that by combining SUMO's high-level description of objects with the output of a machine learning classifier, it is possible to increase in 10 percentage points the precision and recall of such classifier when the scene classification is uncertain. Although this increase is achieved with almost no increase in running time, there as still some scenes that could not be classified for the lack of context regarding the objects found in the image.

We believe that the proposed framework has two compelling applications. The first is to improve statistical classifiers, following an approach similar to (Chen, Shrivastava, and Gupta 2014), where new samples are evaluated before inserting them into the knowledge base. The second is the possibility to include abstractions (in terms of high-level relations, spatial or not) to static data sets in order to enhance image retrieval tasks.

Acknowledgements

Danilo Nunes was partially supported by CAPES (grant 5131024). Leonardo Anjoletto Ferreira acknowledges that this study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001 and also from FAPESP-IBM (grant 2016/18792-9). Paulo Santos acknowledges support from FAPESP-IBM (grant 2016/18792-9).

References

Bloch, I.; Hudelot, C.; and Atif, J. 2007. On the interest of spatial relations and fuzzy representations for ontology-based image interpretation. In *Proceedings of the 7th International Conference on Advances in Pattern Recognition*, ICAPR'07, 15–25. Kolkata, India: ICAPR.

Chen, X.; Shrivastava, A.; and Gupta, A. 2014. Enriching visual knowledge bases via object discovery and segmentation. *Computer Vision and Pattern Recognition*.

Choi, M. J.; Lim, J. J.; Torralba, A.; and Willsky, A. S. 2010. Exploiting hierarchical context on a large database of object categories. In *2010 IEEE Comp. Soc. Conference on Computer Vision and Pattern Recognition*, 129–136.

Csurka, G.; Dance, C. R.; Fan, L.; Willamowski, J.; and Bray, C. 2004. Visual categorization with bags of keypoints. In *Workshop on Statistical Learning in Computer Vision, ECCV*, 1–22.

Dai, B.; Zhang, Y.; and Lin, D. 2017. Detecting visual relationships with deep relational networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Girshick, R.; Donahue, J.; Darrell, T.; and Malik, J. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proc. of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '14, 580–587. Washington, DC, USA: IEEE Computer Society.

Gould, S.; Fulton, R.; and Koller, D. 2009. Decomposing a scene into geometric and semantically consistent regions. In *2009 IEEE 12th International Conference on Computer Vision*, 1–8.

Hare, J. S.; Sinclair, P.; Lewis, P. H.; Martinez, K.; Enser, P. G.; and Sandom, C. J. 2006. Bridging the semantic gap in multimedia information retrieval: Top-down and bottom-up approaches. In *3rd Euro. Semantic Web Conference*, volume 187.

Hudelot, C.; Atif, J.; and Bloch, I. 2008. Fuzzy spatial relation ontology for image interpretation. *Fuzzy Sets Syst.* 159(15):1929–1951.

Jaccard, P. 1912. The distribution of the flora in the alpine zone. 1. *New phytologist* 11(2):37–50.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *Proc. of the 25th International Conference on Neural Information Processing Systems*, NIPS'12, 1097–1105.

Lan, T.; Yang, W.; Wang, Y.; and Mori, G. 2012. Image retrieval with structured object queries using latent ranking svm. In *Proceedings of the 12th European Conference on Computer Vision - Volume Part VI*, ECCV'12, 129–142. Berlin, Heidelberg: Springer-Verlag.

Lehmann, J., and Vöelker, J. 2014. An introduction to ontology learning. In Lehmann, J., and Vöelker, J., eds., *Perspectives on Ontology Learning*. AKA Heidelberg. ix–xvi.

Lu, C.; Krishna, R.; Bernstein, M.; and Fei-Fei, L. 2016. Visual relationship detection with language priors. In *European Conference on Computer Vision*.

Mai, L.; Jin, H.; Lin, Z.; Fang, C.; Brandt, J.; and Liu, F. 2017. Spatial-semantic image search by visual feature synthesis. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Malinowski, M., and Fritz, M. 2014. A pooling approach to modelling spatial relations for image retrieval and annotation. *arXiv:1411.5190 [cs.CV]*.

Niles, I., and Pease, A. 2001. Towards a standard upper ontology. In *Proceedings of the International Conference on Formal Ontology in Information Systems - Volume 2001*, FOIS '01, 2–9. New York, NY, USA: ACM.

Pease, A. 2011. *Ontology: A Practical Guide*. Angwin, CA: Articulate Software Press.

Philbin, J.; Sivic, J.; and Zisserman, A. 2008. Object mining using a matching graph on very large image collections. In *Proceedings of the 2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, ICVGIP '08, 738–745. Washington, DC, USA: IEEE Computer Society.

Sankat, M.; Thakur, R. S.; and Jaloree, S. 2016. *Semi-Automatic Ontology Design for Educational Purposes*. Hershey, PA, USA: IGI Global. 124–142.

Schulz, S. 2013. System Description: E 1.8. In McMillan, K.; Middeldorp, A.; and Voronkov, A., eds., *Proc. of the 19th LPAR, Stellenbosch*, volume 8312 of *LNCS*. Springer.

Siddiquie, B.; White, B.; Sharma, A.; and Davis, L. S. 2014. Multi-modal image retrieval for complex queries using small codes. In *Proc. of International Conference on Multimedia Retrieval*, ICMR '14, 321–328. New York, NY, USA: ACM.

Socher, R.; Lin, C. C.-Y.; Ng, A. Y.; and Manning, C. D. 2011. Parsing natural scenes and natural language with recursive neural networks. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, 129–136. USA: Omnipress.

Wan, J.; Wang, D.; Hoi, S. C. H.; Wu, P.; Zhu, J.; Zhang, Y.; and Li, J. 2014. Deep learning for content-based image retrieval: A comprehensive study. In *Proceedings of the 22Nd ACM International Conference on Multimedia*, MM '14, 157–166. New York, NY, USA: ACM.

Zheng, S.; Jayasumana, S.; Romera-Paredes, B.; Vineet, V.; Su, Z.; Du, D.; Huang, C.; and Torr, P. H. S. 2015. Conditional random fields as recurrent neural networks. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, number 9 in ICCV 15, 1529–1537. Washington, DC, USA: IEEE Computer Society.

Zhou, B.; Zhao, H.; Puig, X.; Fidler, S.; Barriuso, A.; and Torralba, A. 2016. Semantic understanding of scenes through the ade20k dataset. In *arXiv preprint arXiv:1608.05442*.