Spatial descriptions on a functional-geometric spectrum: the location of objects

Simon Dobnik[1][0000−0002−4019−7966] and Mehdi
Ghanimifard[1][0000−0002−2598−5091]

Centre for Linguistic Theory and Studies in Probability (CLASP),
Department of Philosophy, Linguistics and Theory of Science (FLoV),
University of Gothenburg, Box 200, SE 405 30 Gothenburg, Sweden
{simon.dobnik,mehdi.ghanimifard}@gu.se

Abstract. Experimental research on spatial descriptions shows that their semantics are dependent on several modalities, among others (i) a geometric representation of space ("where", geometric knowledge) and (ii) dynamic kinematic routines between objects that are related ("what", functional knowledge). In this paper we examine whether geometric and functional bias of spatial relations is also reflected in large corpora of images and their corresponding descriptions. In particular, we examine whether the variation in object locations in the usage of a relation is a predictor of that relation's functional or geometric bias. Previous experimental psycho-linguistic work has examined the bias of some spatial relations, however our corpus-based computational analysis allows us to examine the bias of spatial relations and verbs beyond those that have been tested experimentally. Our findings have also implications for building computational image descriptions systems as we demonstrate what kind of representational knowledge is required to model spatial relations contained in them.

Keywords: spatial descriptions · geometric · functional · corpus · image captions · computational model

1 Introduction

The work on spatial relations such as "the chair is to the left of the table" and "the bicycle near the door" shows that the semantics of spatial relations is complex, drawing on several different modalities which include among others (i) scene geometry, (ii) functional interactions between objects, and (iii) dialogue interaction between conversational partners. For example, [19] argue that language encodes objects and places differently and this may be a reflection of different cognitive processes in the visual system: "what" and "where". Further, a number of papers [7,6,5,14] show experimentally that different spatial relations have different bias in terms of functional ("what") and geometric ("where") knowledge. Similarly, [18] argues that two classes of spatial relations have different developmental trajectories and may be rooted in different neural representations. [8] argues that the bias to function and geometry of a particular relation

is contextual and task-dependent. It is important to note that, since objects are grounded in space, their functional properties and interaction between them are also reflected in their geometric representations, in particular how they are conceptualised as different geometric shapes and how they are arranged in scene configurations (cf. [13]). However, the background knowledge that allows us to do this geometric projection of scenes comes from our conceptual understanding of the world, for example our knowledge that bowls are used to contain fruit and umbrellas are used to protect people from the rain.

For this reason, computational modelling of descriptions of spatial relations is challenging. Firstly, it requires information from each of these modalities to be present in the dataset. For example, it is hard to collect a large enough dataset of functional interactions between objects and represent these interactions as computationally useful representations. Secondly, there is a challenge of information fusion which needs to be attuned for different words in different contexts. Recently, deep neural networks modelling language and vision as perceptually grounded language models have demonstrated a lot of success [28,21]. An interesting research question therefore is what information such networks can capture in their representations from the available modalities and whether such representations correspond to the representations that have been argued for in linguistic and psychological literature.

For example, [10,11,9] explore whether functional and geometric bias can be recovered from the information encoded in a language model, the semantic associations encoded in the sequences of words. Language models together with word embeddings [2] are widely used to represent linguistic meaning in computational semantics and they are based on the premise known as the *distributional hypothesis* [12] that words occuring in similar contexts, represented by other words, will have similar meanings [27]. If we relate the distributional hypothesis to grounding in perception, this is because words co-occuring together will refer to identical situations and therefore the contexts of words become proxies for accessing the underlying situations. It follows that information encoded in language models about spatial descriptions should encode some relevant semantics about dynamic kinematic routines between the objects that are related, albeit very indirectly. Hence, [10,11,9] demonstrate that the functional-geometric bias of expressions that have been tested experimentally in [7] is reflected in the degree to which target and landmark objects are associated with a relation in spatial descriptions extracted from a corpus of image descriptions. They start with the idea that while any two (abstract) objects can be related in geometric space, functional relations between the objects and relation are more specific, defined by the possible functional interaction between the objects. They demonstrate that this is expressed in the variability and generality of the target and landmark objects. Since a geometrically-biased spatial relation can relate any kind of objects that can be placed in a particular space, the objects used with such a relation will be more variable than the objects that occur with functionally-biased relations that also encode the nature of object interaction. They also show that usage of descriptions of an image corpus is crucial in this task since in a general corpus,

a wider range of situations is reflected in the word contexts that may include metaphoric usages of the spatial words in other domains that do not involve spatial geometry. We may consider such metaphorical usage of spatial relations in other domains as highly functional.

The experiments based on [7] show that spatial relations have functional or geometric *bias* which means that both components are relevant for the semantics of a description, just not the same degree. For example, a functionally-biased relation such as *over* is also sensitive to geometry to some extent, it appears that a presence of a function skews the regions of acceptability for the target object of that relation. The deviation in geometry can be explained by the fact that under a consideration of a functional relation different parts of the target and landmark object will become attended [5,3]. This results in a situation where the centroids of bounding boxes of target and landmark objects are displaced from the locations where we would expect to find them based on the geometric constraints alone. For example, in the case of a "teapot over a cup" it must be ensured that the spout of the teapot is located in such a way so that the liquid will be poured into a cup. In a scene described by a description "the toothpaste is over a toothbrush" the shape of the bounding boxes will be different from the previous scene as well as the location of the attended areas. In the case of an "apple in a bowl" the bowl or its contents must constrain the movement of the apple (so that it does not fall out of the bowl) and hence locations of apples that are outside the bounding box of the bowl are also acceptable, for example where an apple is on the top of other apples. These examples suggest that over all contexts of target-landmark objects, the variation in locations of objects represented as bounding boxes will be much higher with functionally-biased spatial relations than geometrically-biased ones which will be closer to the axes of the geometric space. The latter is confirmed by the spatial templates of [20] where in the absence of the functional knowledge, when an abstract shapes are used as targets and landmarks, both geometric and functional relations such as "over" and "above" give very similar axis-centred spatial templates. Hence, in this work, we explore whether we can detect a difference in the variability of the target objects in relation to the landmark objects for spatial relations of either geometric or functional bias in terms of representations of objects as visual features in images from a large corpus of images and descriptions and for relations that go beyond the ones that were tested experimentally. We expect that this variability will be the opposite of the variability that has been previously shown for textual data [10,11,9]. Functional information can be recovered from the textual information about *what* objects are interacting, while geometric information can be recovered from *where* the visual features of objects are. Hence, we expect that relations that were experimentally found to have a functional bias will be less variable in their choice of target and landmark objects but more variable in terms of where these objects are in relation to the prototypical axes from the landmark. On the other hand, relations that were experimentally found to have a geometric bias, are expected show a higher variation in terms of the object kinds they relate but these will be geometrically less variable from the axes based on the landmark.

The experimental work on functional and geometric bias of spatial relations focuses on abstract images where the type of objects, their location and the nature of functional interaction is carefully controlled. This gives us accurate judgements about the applicability of descriptions but since the task focuses on abstract scenes this gives us different judgements to those we would have hoped to have obtained in real-life situations simply because of the perceptual and linguistic context is different from real-life situations [8]. Ideally, we would need a corpus of interactions between real objects and their spatial descriptions that on the perceptual side would be represented as 3-dimensional temporal model. Collecting such a corpus on a large scale would be a very challenging endeavour, although important work in this area has recently been done in route instructions in a virtual environments [26]. On the other hand, there exist several large corpora of image descriptions, e.g. [16] which contain spatial descriptions and a large variety of interacting objects in real-life situations. For this reason they are, in our opinion, an attractive test-bed for examining the meaning of geometrically-biased and functionally-biased spatial relations. The down-side of image corpora is that the visual representations scenes are skewed, depending on the angle and the focus/scale at which an image was taken which means that an object such as a chair may have a different shape and size in respect to the image from one image to another. There is also no information about object depth and the dynamic interaction of objects. To counter this variation in objects we will introduce some normalisation steps. Of course, there will also be some noise in the scene representation's we obtain but we hope this noise will be uniform across different images and kinds of descriptions and therefore a relative comparison of descriptions of different bias will still give us a valid result.

Why is identification of functional and geometric bias of spatial relations relevant? Theoretically, the experiments give us more insights into the way spatial cognition is reflected in language. Showing that there is a distinction between these two classes of spatial relations on a large scale dataset of image descriptions gives a further support to the experimental evidence that has been obtained in carefully designed experiments. Knowing that there are different classes of spatial relations can help us in the task of generating image descriptions, for example in a robotic scenario. Following our observation, in an image description task functional relations are more informative than geometric relations as in addition to geometric component they also say something about the relation between the objects.[1] In a given scene a target object can be described and related to the landmark with several spatial relations based on geometric considerations alone. However, these descriptions could be filtered by considering those relations that are functionally more likely. The investigation also has implication for end-to-end image captioning systems build with deep learning architectures. Knowing

---

[1] Notice, however, that there are tasks where geometric information may be more informative, for example when answering a question about the location of an object in a visual scene. The choice of a spatial relation therefore depends on the communicative intent of the speaker and the task they are engaged in.

that different spatial relations have a different bias for visual and textual modality would allow us a better comparison and evaluation of such systems. For example, there is a significant discussion in the vision and language community that end-to-end image captioning systems and visual question answering systems are relying too much on the information from language models [1] rather than grounding words in an image, particularly when it comes to describing relations between objects. Knowing that not all spatial relations are equally geometrically spatial has important implications for evaluating such systems: (i) it shows that provided there is a balanced dataset reliance of a spatial relation on a language model is not necessarily a shortcoming but rather that is in fact the dimension that determines their meaning and there is a gradience in the way a description is grounded in visual vs textual features; (ii) it gives us insights into how we should build such systems in the future so that both (or even more) modalities are appropriately represented.

This paper is organised as follows: in Section 2 we describe the dataset of images and descriptions used in our studies; in Section 3 we describe how we represent geometric information from image annotations for spatial relations and how such representations can be compared for functional and geometric bias; in Section 4 we introduce a more sophisticated comparison in terms of the variation in our feature representations for different spatial relations from a representative representation; and we conclude in Section 5.

2   Dataset

We base our investigations on the Visual Genome dataset [16] which is a crowd-sourced annotations of 108,007 images. The dataset comprises several types of annotations including the region descriptions (phrases and sentences referring to one bounding box), objects (annotated as bounding boxes), attributes for each object annotation, and *relationships* between them (triplet of subject, predicate, object). Most object names, attributes and predicate of relationships are also mapped to WordNet synsets. The predicates in relationships include *spatial relations* such as "*above*", "*under*", "*on*", "*in*" but also verbs describing events such as "*holding*" and "*wearing*", or a combination of both such as "*sitting on*".

Without any data cleaning, the total number of possible forms of relation tokens is 36,550. Since spatial relations are multi-word expressions, we create a dictionary of relations capturing different variations of their syntactic form (e.g. "to the left of", "on the left", "left", etc.) based on the lists of English spatial relation constructions in [17] and [13]. Out of 235 spatial relations, we only found 78 types. Some variation in writing of relationships may be simply due to the annotator shorthand notation, e.g. "to left of". We combine the compound variants of spatial relations to a lower-cased single variant in cases where we can be reasonably sure that this will not affect their semantics in terms of functional and geometric bias. Duplicate descriptions per image which are created by different annotators are removed, as well as those descriptions where the extracted spatial relations are not used in a complete locative description

involving a target object, relation and a landmark, e.g. "chair on left". At the end, we only kept those relations which have more than 30 instances in the dataset.

In addition to spatial relations, we also added a few verbal relations that describe situations that are grounded in space, for example verbs that [4] have shown to have strong predictability of object on the y-axis. The dictionary of all relations examined in this study is given in Table 1.

Table 1. The list of consolidated spatial relations and verbs.

over, above, below, under, left of, right of, on, in, inside, outside, far from, away from, next to, near to, across, at, with, beneath, underneath, through, alongside, against, off, between, from, beside, to, by, along, around, behind, bottom, top, front of, back of, side of,
flying, kicking, cutting, catching, riding, seeing, looking, floating, finding, pulling, removing, having, wearing, containing, holding, supporting, sitting, touching.

3   Representing locations as dense geometric vectors

Each bounding box in Visual Genome is represented with 4 numerical values: the x-, y- coordinates relative to the image frame, the bounding box width and height. In order to compare the geometric arrangements of objects represented as bounding boxes between different spatial relations, as well as to compare this data with the data from spatial templates from [20], we convert both representations to 3-dimensional dense vectors $[x, y, d]$ where x and y represent directions in the 2-dimensional space and d is a Euclidean distance between x and y. Hence, we separate directionality (represented by x and y) from the distance. The intuition behind this comes from a distinction between *projective relations* ("to the left of" and "above") and *topological relations* ("in", "at", "near") where the former are dependent on both directionality and distance but the latter are only dependent on distance. The 3-dimensional vectors (the x and y dimension) are inspired by vectors introduced in the Attentional Vector Sum Model (AVS) [23]. However, as we will describe below they are used quite differently. Rather then modelling the attention for a particular pair of bounding boxes in the AVS model we use them to estimate attention between all bounding boxes that are related by a particular spatial relation. In other words, we use them to estimate the likelihood that for a particular spatial relation a particular location is occupied by an object. Therefore, the representations are similar to the notion of spatial templates. Here, other representations of bounding boxes could also be used (see for example [24,22]. No doubt, different geometric representations favour different classes of spatial relations differently and this will be reflected in our results. For example, we expect that our 3-dimensional dense vectors are not suited to ground relations such as "around" that require grounding in multiple

locations at different sides. This raises two interesting questions that have no straightforward answers: what are basic geometric representations required to model spatial language and to what degree is the choice of what representations go into our geometric framework a part of the functional knowledge. Overall, we opt for simplest low-level geometric representations that are used in spatial templates and the AVS model.
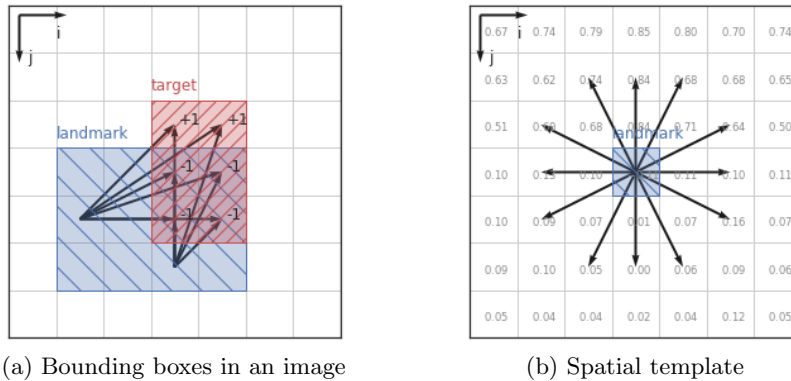

(a) Bounding boxes in an image   (b) Spatial template

Fig. 1. (a) Images are segmented to a fixed set of locations and relation vectors are calculated for every pair of locations occupied by the bounding boxes of target and landmark. (b) In spatial templates a vector is calculated for every location of the template originating in the location of the landmark.

We derive the dense features as follows. First, as shown in Figure 1a, we segment images into 7×7 locations. Then, for every pair of points in the locations matrix, we define a dense vector as:

$$\text{for two points on image} \begin{cases} p_1 = \langle i_1, j_1 \rangle \\ p_2 = \langle i_2, j_2 \rangle \end{cases}, \boldsymbol{u}_{p_1, p_2} = \begin{bmatrix} x \\ y \\ d \end{bmatrix} = \begin{bmatrix} \frac{i_2 - i_1}{||\overrightarrow{p_1 p_2}||_2} \\ \frac{j_1 - j_2}{||\overrightarrow{p_1 p_2}||_2} \\ \texttt{sgn} \cdot ||\overrightarrow{p_1 p_2}||_2 \end{bmatrix}$$

where $\boldsymbol{u}_{p_1, p_2}$ represents the dense geometric relation features between two points, which p_1 is a point on landmark and p_2 is on the target, the Euclidean distance between them is $||\overrightarrow{p_1 p_2}||_2 = \sqrt{(i_2 - i_1)^2 + (j_2 - j_1)^2}$, and $\texttt{sgn}$ is a sign value which is -1 if p_2 is also a point on the landmark bounding box, otherwise $+1$.

For each relation REL, this gives us a collection of vectors. For bounding boxes annotated with relations in the images of Visual Genome, we build the collection of dense vectors of all points connecting targets and landmarks related by each particular relation in the dataset ($V_{\text{REL}}^{(vg)}$). Formally, this set is represented as

follows:

$$V_{\text{REL}}^{(vg)} = \left\{ \boldsymbol{u}_{p_1,p_2} \right\}_{\substack{\langle \text{TRG,REL,LND} \rangle \in \text{Images} \\ p_1 \in \text{bbox}_{\text{LND}} \\ p_2 \in \text{bbox}_{\text{TRG}}}} \tag{1}$$

where bbox_{TRG} and bbox_{LND} are the collection of points in bounding boxes of target TRG and landmark LND.[2]

Similarly, we use this method on spatial templates from [20] to build all possible dense vectors. As shown in Figure 1b, we create a dense vector originating in the central location of the landmark and ending at every possible location of target in the spatial template. Each vector from a spatial template is associated with the acceptability score of the target location.

$$V^{(st)} = \left\{ \boldsymbol{u}_{\langle 3,3 \rangle, \langle i,j \rangle} \right\}_{\substack{i \in \{1,...,7\} \\ j \in \{1,...,7\}}}, \quad S_{\text{REL}} = \left\{ \boldsymbol{s}_{i,j} \right\}_{\substack{i \in \{1,...,7\} \\ j \in \{1,...,7\}}} \tag{2}$$

where S_{REL} represents the collection of normalised acceptabilities in spatial template of the relation REL.

These vectors in each collection are then projected to a single vector representation using the following methods. For the collection of vectors from a spatial template, the representative vector is the weighted sum of all possible vectors with acceptability scores:

$$\boldsymbol{v}_{\text{REL}}^{(st)} = \sum_{\substack{i \in \{1,...,7\} \\ j \in \{1,...,7\}}} s_{i,j} \cdot \boldsymbol{u}_{\langle 3,3 \rangle, \langle i,j \rangle} \tag{3}$$

For the collection of vectors from the Visual Genome bounding boxes, the representative vector is the expected 3-feature vector:

$$\boldsymbol{v}_{\text{REL}}^{(vg)} = E[V_{\text{REL}}^{(vg)}] = \frac{1}{|V_{\text{REL}}^{(vg)}|} \sum_{\boldsymbol{v} \in V_{\text{REL}}^{(vg)}} \boldsymbol{v} \tag{4}$$

where $|V_{\text{REL}}^{(vg)}|$ is the number of vectors. Adding vectors with contradicting features will cancel each other and the remaining vector points at a direction with the least opposite directions. More importantly, the resulting three dimensional feature vector $\boldsymbol{v}_{\text{REL}}^{(vg)}$ from bounding box annotations in Visual Genome is similar to $\boldsymbol{v}_{\text{REL}}^{(st)}$ from spatial templates.

To compare the projected dense vectors we have obtained from the images with those from the spatial templates we use cosine similarity or distance as shown in Figure 3 where the horizontal axis represents the vectors from spatial templates $\boldsymbol{v}_{\text{REL}}^{(st)}$ and the vertical axis represents the vectors from images $\boldsymbol{v}_{\text{REL}}^{(vg)}$. The results indicate that the 3-dimensional vectors from the two datasets are very

---

[2] For computational convenience, instead of including all possible annotations in this set, we randomly sampled a maximum of 1000 triplets from the relationship dataset.
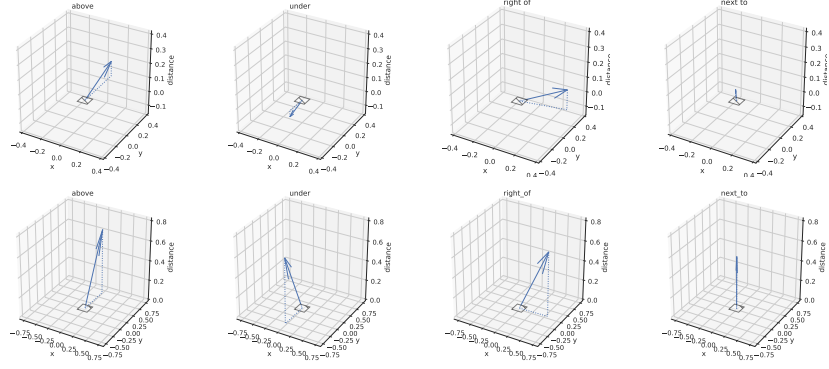
Fig. 2. Examples of $\boldsymbol{v}_{\text{REL}}^{(vg)}$ and $\boldsymbol{v}_{\text{REL}}^{(st)}$: vectors are similar in all three dimensions but their origin and scale are different.
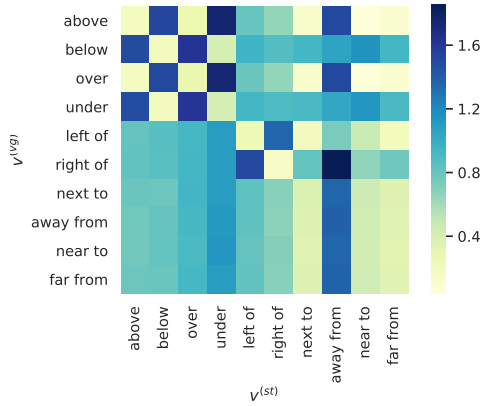

Fig. 3. A comparison of dense vector representations from images $\boldsymbol{v}_{\text{REL}}^{(vg)}$ and those from spatial templates $\boldsymbol{v}_{\text{REL}}^{(st)}$ with the cosine distance: $1 - cosine(\boldsymbol{v}_{\text{REL}}^{(vg)}, \boldsymbol{v}_{\text{REL}}^{(st)})$.

similar except in the case of "away from". Except for this case the lowest cosine distance is on the diagonal. The results also indicate that pairs of geometrically or functionally biased spatial relations such as "over" and "above" and "under" and "below" have similar overall directions and distances. Projective relations have clearly defined opposites alongside one axis but topological relations are overlapping with the projective relations. "next to$_{st}$" is similar to "next to$_{vg}$", "away from$_{vg}$", "near to$_{vg}$" and "far from$_{vg}$" and "away from$_{st}$" is dissimilar to all. This has possibly to do with the way distance is represented in images. Humans are able to estimate distance between two focused objects not on their actual size but the size they know from their background knowledge.

The comparison of dense vectors here indicates that similar dense vectors are obtained from both datasets. However, it does not distinguish functional and geometric bias of different relations. For example, "over$_{st}$" is equally similar to "over$_{vg}$" and "above$_{vg}$" while we were expecting that since "over$_{st}$" is used in the geometric context it will be similar to "above$_{vg}$". This is because cosine similarity/distance takes into account all three dimensions x, y and z of the dense vectors. However, we expect that "over$_{st}$" will be similar to "over$_{vg}$" in y and d dimensions but different in the x dimension which distinguishes its geometric and functional use.

In the following section we examine the 3-dimensional feature space of the dense vectors in terms of the variation in the distribution of features. Therefore, we need to look for a measure that captures variation in distribution of features.

4   Variation of features within dense vectors

We argued in Section 1 that we expect that functionally-biased relations will be associated with more variable locations of target and landmark objects as these will also be dependent on the functional relations between individual object pairs. In the previous section we represented the locations between targets and landmarks as dense vectors which were then projected to one representative vector for each spatial relation. The degree of divergence from the representative vectors can be considered as an indication for non-geometrical use of spatial relations. In order to test this, for each spatial relation, we calculate a deviation of individual target-landmark vectors $\boldsymbol{v}$ from the representative 3-dimensional dense vector $\boldsymbol{v}_{\mathrm{REL}}^{(vg)}$. As a metric of deviation we use cosine distance:

$$Distances = \left\{ 1 - cosine(\boldsymbol{v}_{\mathrm{REL}}^{(vg)}, \boldsymbol{v}) \right\}_{\boldsymbol{v} \in V_{\mathrm{REL}}^{(vg)}} \tag{5}$$

We expect that on average, cosine distances in geometrically-bias relations are closer to 0 (there is a clearer central tendency), and the overall distribution of cosine distances is positively skewed: the mode of cosine distances is close to zero while the mean and the tail of differences is skewed to the right.[3]

---

[3] To calculate skewness we use an implementation of the Fisher-Pearson coefficient [15, s.2.2.24.1] in `scipy.stats.skew`.
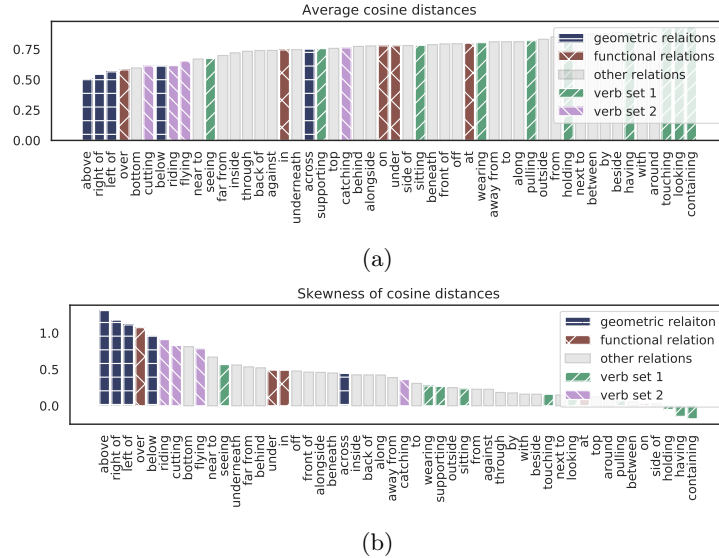
(a)

(b)

Fig. 4. (a) The average cosine distance of dense vectors $[x, y, d]$ from the expected dense vector of each spatial relation. (b) The skewness in distribution of distances. The colour indicates the status of each relation as reported in the literature.

In Figure 4, we select a set of geometrically- and functionally-biased relations that have been experimentally tested and reported in the literature and plot (a) their average cosine distances of dense vectors from their representative vector and (b) the skewness of cosine differences. We also include relations the bias of which has not been tested experimentally (other) but we expect that this is demonstrated by their position in the graph between the key-points determined experimentally. Finally, we also include some verbs describing events and situations involving interacting objects in space that are also annotated as relationships in the Visual Genome, e.g. "boy, feeds, giraffe". We are particularly interested in the verbs that are reported in Collell et al. [4] for which the location of the (target) object is most strongly predictable from the y dimension ("flying", "kicking", "cutting", "catching" and "riding") (verb set 2 in Figure 4) and those for which the y dimension is the least predictable in respect to the location of the object ("see", "float", "finding", "pulled" and "removes") (verb set 1) listed in their Table 3, p.6770. Note, however, that [4] do not consider the x-dimension which may be a relevant dimension for the verbs in the picture. A quick comparison of the two lists gives an impression that the former contains descriptions of events involving object relations that are more strongly grounded in the image representations (e.g. "riding") and are therefore similar to geometrically-biased spatial relations, while the second list contains descriptions of events that are less strongly grounded in the image representations (e.g.

"sees") and would require a simulation of dynamic kinematic routines between the objects which makes them similar to functionally-biased spatial relations.

Examining the average cosine distances from the representation vector of each spatial relation in Figure 4a we can see that relations that have been identified as geometrically-biased tend to have a lower average cosine distance from the representation's dense vector than those that have been identified as functionally-biased. The same tends also to be the case for verbs identified in [4] for which the objects are more dependent on the y (verb set 2) compared to verbs for which the objects are less dependent on the y dimension (verb set 1). Note that in this comparison a deviation of the entire 3-dimensional vector $[x, y, d]$ was taken into account and therefore a deviation can be in any of these dimensions. Examining the skewness of cosine distances from the representation vector of each spatial relation in Figure 4b we can see that geometrically-biased verbs and verbs that are more strongly grounded show a tendency towards a higher skewness of distribution, they are more biased towards the representational vectors. Overall, the results indicate support for our hypothesis in Section 1 that bounding boxes are predictors of the functional and geometric bias as well as they indicate that the same bias is also present in verbal descriptions of scenes.

In Figure 5 we examine the histograms of deviations from the representational vectors of "on", "in", "over", "above", "right of" and "left of". To plot these histograms we use Kernel Density Estimation (KDE)[4] [25] which indicates the density of samples in the range of $[0, 2]$ of the cosine distance (Equation 5). We also give examples of target-landmark pairs which have the highest (brown/light) and the lowest (blue/dark) average distances from the representational vectors. These examples indicate that functionally biased relations ("on", "in" and "over") are used in contexts where the geometric constraint is satisfied and also in contexts where there is a deviation from the geometric constraint, just as predicted by experiments in [7]. Interestingly, among the cases that show high deviation from the representational vectors we also find examples that are typically considered to involve more complex geometric conceptualisations which we argued are a result of taking into account object function, for example "bracelet on wrist", "woman in dress", and "trees over rocks". However, within the relations that we consider to be geometrically-biased we also find examples of high deviation from the representational vectors. The examples for "above" correspond to usages where there is an element of covering or protection that has been argued to be the functional component of "over", e.g. "clouds above/over pasture" and "mirror above/over bench" or cases that require complex geometric conceptualisation of the scene, e.g. "tree above ground". We are intrigued by the examples that deviate from the representational vectors for "left of" and "right of". They frequently contain animate beings or objects with clearly defined front and back and therefore have orientation. Our interpretation is that these examples are a reflection of changes of the perspective from the relative frame of reference of the observer of the image to the intrinsic frame of reference of the landmark. Since our framework assumes the relative frame of reference by default, the change to

---

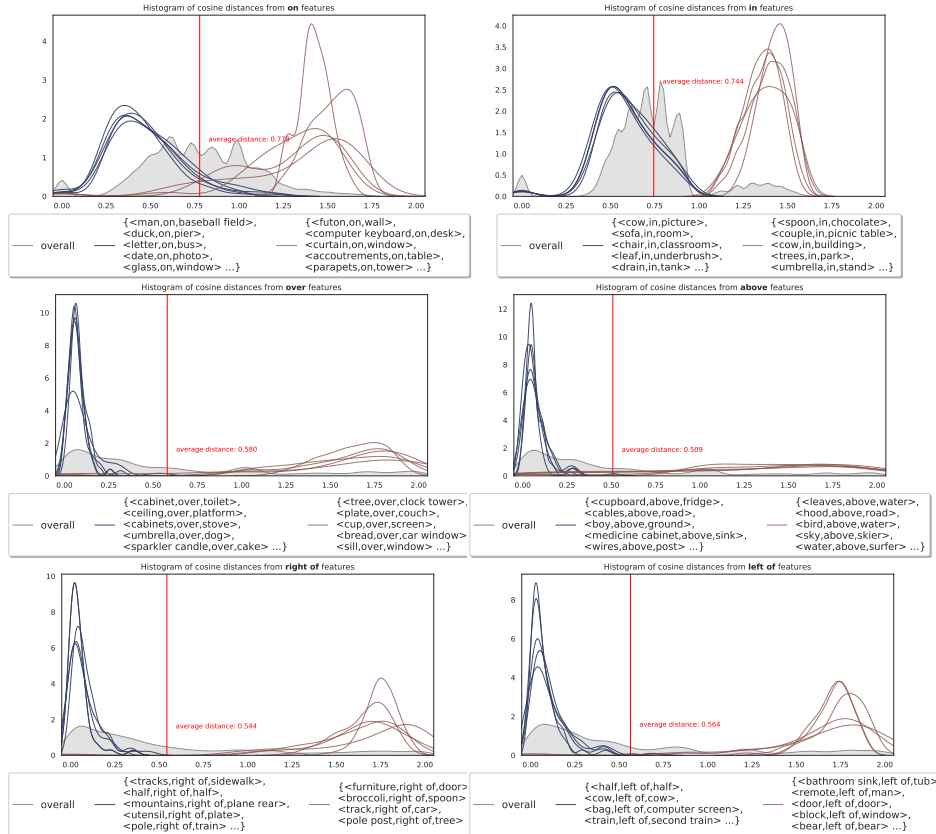[4] We use an implementation based on `scipy.stats.gaussian_kde`

Fig. 5. Using the KDE method we plot a histogram of cosine distances of individual examples from the representational vector of each relation. The histogram shows skewness to zero for geometrically-biased usages of relations. The individual lines show some examples of target-landmark pairs which have the lowest (blue/dark) and the highest (brown/light) average distance from the representational vectors.

a different frame of reference in a description would lead to high distances in our results.
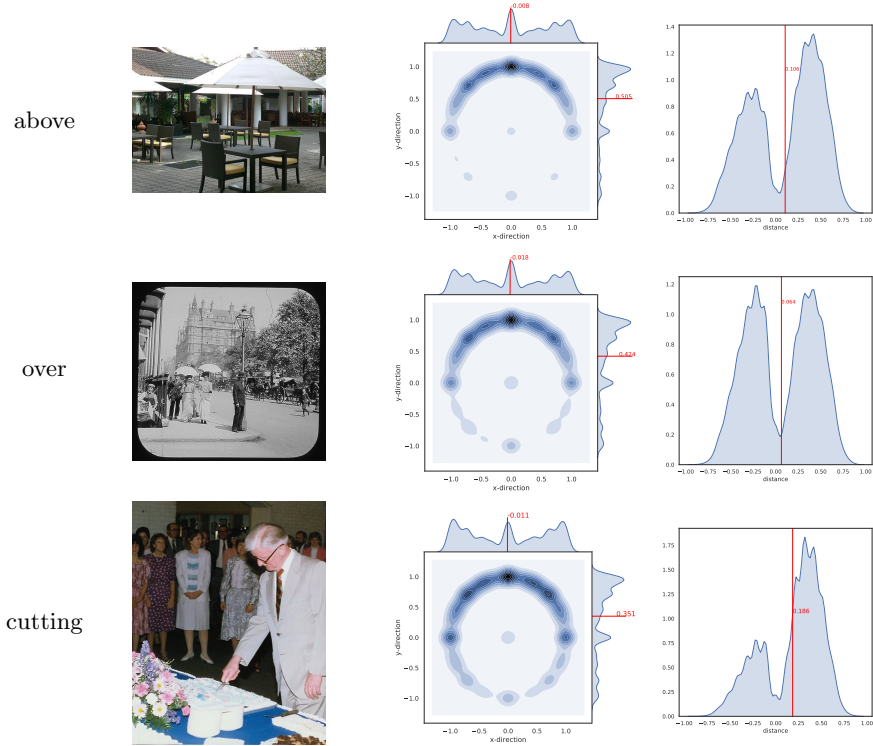

Fig. 6. The individual features of dense vectors $[x, y, d]$ have different distributions for different relations.

As stated earlier, the dense vector representations including their cosine distances aggregate three features $[x, y, d]$ and therefore the previous comparisons do not take into account the role of each individual feature for spatial relations. In Figure 6 we plot the distribution of all features over all vectors of $V_{\mathrm{REL}}^{(vg)}$ for some relations. These relations were found to be strongly dependent on the y feature in [4] who only considered this feature. The individual histograms for the x (centre top), y (centre right) and d feature (on the right side) indicate the density of their values and the mixture density graph for x, y (centre) shows how these features interact. This graph demonstrates that "over" and "cutting" have more freedom of variation in the x dimension as well as the negative y dimension (which indicates overlap of objects) than "above". As discussed earlier, there is also considerable overlap between all three graphs which is due to the fact that functionally-biased relations are also used in situations when geometric

constraints are satisfied. While "cutting" is more similar to "over" than "above" in terms of the xy dimensions, it has a very different distance dimension.

5   Conclusion

In this paper we have demonstrated how the functional and geometric bias of spatial relations can be identified from geometric annotations of objects as bounding boxes connected by spatial relations in a corpus of images and associated descriptions. The bounding boxes are converted to 3-dimensional dense vectors that contain information about the x, y and d dimensions. These vectors are then merged to a single representational vector for each spatial relation. Vectors for different relations are then compared with cosine similarity. To increase the granularity of comparison we examine how individual examples of annotated situations diverge from the representational vectors and what are the distributions of these divergences, also at the level of individual features. Our results indicate that functional and geometric bias of spatial relations can be identified from the geometric spatial information captured in a large corpus of images and descriptions and that this distinction can be carried over to verbs describing situations involving objects. Our study makes a contribution to the study of semantics of spatial descriptions by demonstrating that information that was previously determined experimentally under constrained conditions for a smaller number of spatial relations can be replicated on a larger scale and in noisy contexts. Practically, such information is extremely useful for building end-to-end deep neural models of image captioning as it demonstrates what kind of representations are relevant for different kinds of descriptions which has also been the focus of our other studies. Another question that we find relevant to explore in our future work is the observation that the context in which the dataset was created introduces a general bias on the degree to which function and geometry is considered to be relevant. For example, is the intent of the image description task to describe *what* is happening to the objects in the scene or to locate *where* the objects are. Finally, different classes of verbs would also deserve a more focused study.

Acknowledgements

References

1. Agrawal, A., Batra, D., Parikh, D., Kembhavi, A.: Don't just assume; look and answer: Overcoming priors for visual question answering. In: Proceedings of the

IEEE Conference on Computer Vision and Pattern Recognition. pp. 4971–4980 (2018)

2. Bengio, Y., Ducharme, R., Vincent, P., Janvin, C.: A neural probabilistic language model. Journal of Machine Learning Research **3**(6), 1137–1155 (2003)

3. Carlson, L.A., Regier, T., Lopez, W., Corrigan, B.: Attention unites form and function in spatial language. Spatial Cognition & Computation **6**(4), 295–308 (2006)

4. Collell, G., Van Gool, L., Moens, M.F.: Acquiring common sense spatial knowledge through implicit spatial templates. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)

5. Coventry, K.R., Cangelosi, A., Rajapakse, R., Bacon, A., Newstead, S., Joyce, D., Richards, L.V.: Spatial prepositions and vague quantifiers: Implementing the functional geometric framework. In: Freksa, C., Knauff, M., Krieg-Brückner, B., Nebel, B., Barkowsky, T. (eds.) Spatial Cognition IV. Reasoning, Action, Interaction, Lecture Notes in Computer Science, vol. 3343, pp. 98–110. Springer Berlin Heidelberg (2005)

6. Coventry, K.R., Garrod, S.C.: Saying, seeing, and acting: the psychological semantics of spatial prepositions. Psychology Press, Hove, East Sussex (2004)

7. Coventry, K.R., Prat-Sala, M., Richards, L.: The interplay between geometry and function in the apprehension of Over, Under, Above and Below. Journal of Memory and Language **44**(3), 376–398 (2001)

8. Dobnik, S., Åstbom, A.: (Perceptual) grounding as interaction. In: Petukhova, V., Tian, Y. (eds.) Proceedings of Saardial – Semdial 2017: The 21st Workshop on the Semantics and Pragmatics of Dialogue. pp. 17–26. Saarbrücken, Germany (August 15–17 2017)

9. Dobnik, S., Ghanimifard, M., Kelleher, J.: Exploring the functional and geometric bias of spatial relations using neural language models. In: Proceedings of the First International Workshop on Spatial Language Understanding. pp. 1–11. Association for Computational Linguistics, New Orleans (Jun 2018)

10. Dobnik, S., Kelleher, J.D.: Towards an automatic identification of functional and geometric spatial prepositions. In: Proceedings of PRE-CogSsci 2013: Production of referring expressions – bridging the gap between cognitive and computational approaches to reference. pp. 1–6. Berlin, Germany (31 July 2013)

11. Dobnik, S., Kelleher, J.D.: Exploration of functional semantics of prepositions from corpora of descriptions of visual scenes. In: Proceedings of the Third V&L Net Workshop on Vision and Language. pp. 33–37. Dublin City University and the Association for Computational Linguistics, Dublin, Ireland (August 2014)

12. Firth, J.R.: A synopsis of linguistic theory 1930–1955. Studies in linguistic analysis pp. 1–32 (1957)

13. Herskovits, A.: Language and spatial cognition: an interdisciplinary study of the prepositions in English. Cambridge University Press, Cambridge (1986)

14. Hörberg, T.: Influences of form and function on the acceptability of projective prepositions in Swedish. Spatial Cognition & Computation **8**(3), 193–218 (2008)

15. Kokoska, S., Zwillinger, D.: CRC standard probability and statistics tables and formulae. Crc Press (2000)

16. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., Bernstein, M., Fei-Fei, L.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. International Journal of Computer Vision **123**(1), 32–73 (May 2017)

17. Landau, B.: Multiple geometric representations of objects in languages and language learners. Language and space pp. 317–363 (1996)

18. Landau, B.: Update on "what" and "where" in spatial language: A new division of labor for spatial terms. Cognitive Science **41**(2), 321–350 (2016)
19. Landau, B., Jackendoff, R.: "what" and "where" in spatial language and spatial cognition. Behavioral and Brain Sciences **16**(2), 217–238, 255–265 (1993)
20. Logan, G.D., Sadler, D.D.: A computational analysis of the apprehension of spatial relations. In: Bloom, P., Peterson, M.A., Nadel, L., Garrett, M.F. (eds.) Language and Space, pp. 493–530. MIT Press, Cambridge, MA (1996)
21. Lu, J., Xiong, C., Parikh, D., Socher, R.: Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 375–383 (2017)
22. Nikolaus, M., Abdou, M., Lamm, M., Aralikatte, R., Elliott, D.: Compositional generalization in image captioning. Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL) (2019)
23. Regier, T., Carlson, L.A.: Grounding spatial language in perception: an empirical and computational investigation. Journal of Experimental Psychology: General **130**(2), 273–298 (2001)
24. Sadeghi, F., Kumar Divvala, S.K., Farhadi, A.: Viske: Visual knowledge extraction and question answering by visual verification of relation phrases. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1456–1464 (2015)
25. Scott, D.W.: Multivariate density estimation: theory, practice, and visualization. John Wiley & Sons (2015)
26. Thomason, J., Murray, M., Cakmak, M., Zettlemoyer, L.: Vision-and-dialog navigation. In: Conference on Robot Learning (CoRL) (2019)
27. Turney, P.D., Pantel, P., et al.: From frequency to meaning: Vector space models of semantics. Journal of artificial intelligence research **37**(1), 141–188 (2010)
28. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: International conference on machine learning. pp. 2048–2057 (2015)