Query-by-Example Image Retrieval using Visual Dependency Representations

OPEN ACCESS

Query-by-Example Image Retrieval
using Visual Dependency Representations

Desmond Elliott, Victor Lavrenko and **Frank Keller**
Institute of Language, Communication, and Computation
School of Informatics
University of Edinburgh
`d.elliott@ed.ac.uk` {`vlavrenk,keller`}`@inf.ed.ac.uk`

Abstract

Image retrieval models typically represent images as bags-of-terms, a representation that is well-suited to matching images based on the presence or absence of terms. For some information needs, such as searching for images of people performing actions, it may be useful to retain data about how parts of an image relate to each other. If the underlying representation of an image can distinguish between images where objects only co-occur from images where people are interacting with objects, then it should be possible to improve retrieval performance. In this paper we model the spatial relationships between image regions using Visual Dependency Representations, a structured image representation that makes it possible to distinguish between object co-occurrence and interaction. In a query-by-example image retrieval experiment on data set of people performing actions, we find an 8.8% relative increase in MAP and an 8.6% relative increase in Precision@10 when images are represented using the Visual Dependency Representation compared to a bag-of-terms baseline.

1 Introduction

Every day millions of people search for images on the web, both professionally and for personal amusement. The majority of image searches are aimed at finding a particular named entity, such as *Justin Bieber* or *supernova*, and a typical image retrieval system is well-suited to this type of information need because it represents an image as a bag-of-terms drawn from data surrounding the image, such as text, manual tags, and anchor text (Datta et al., 2008). It is not always possible to find useful terms in the surrounding data; the last decade has seen advances in automatic methods for assigning terms to images that have neither user-assigned tags, nor a textual description (Duygulu et al., 2002; Lavrenko et al., 2003; Guillaumin and Mensink, 2009). These automatic methods learn to associate the presence and absence of labels with the visual characteristics of an image, such as colour and texture distributions, shape, and points of interest, and can automatically generate a bag of terms for an unlabelled image.

It is important to remember that not all information needs are entity-based: people also search for images reflecting a mood, such as *people having fun at a party*, or an action, such as *using a computer*. The bag-of-terms representation is limited to matching images based on the *presence or absence* of terms, and not the *relation* of the terms to each other. Figures 1(a) and (b) highlight the problem with using unstructured representations for image retrieval: there is a person and a computer in both images but only (a) depicts a person actually using the computer. To address this problem with unstructured representations we propose to represent the structure of an image using the Visual Dependency Representation (Elliott and Keller, 2013). The Visual Dependency Representation is a directed labelled graph over the regions of an image that captures the spatial relationships between regions. The representation is inspired by evidence from the psychology literature that people are better at recognising and searching for objects when the spatial relationships between the objects in the image are consistent with our expectations of the world.(Biederman, 1972; Bar and Ullman, 1996). In an automatic image description task, Elliott
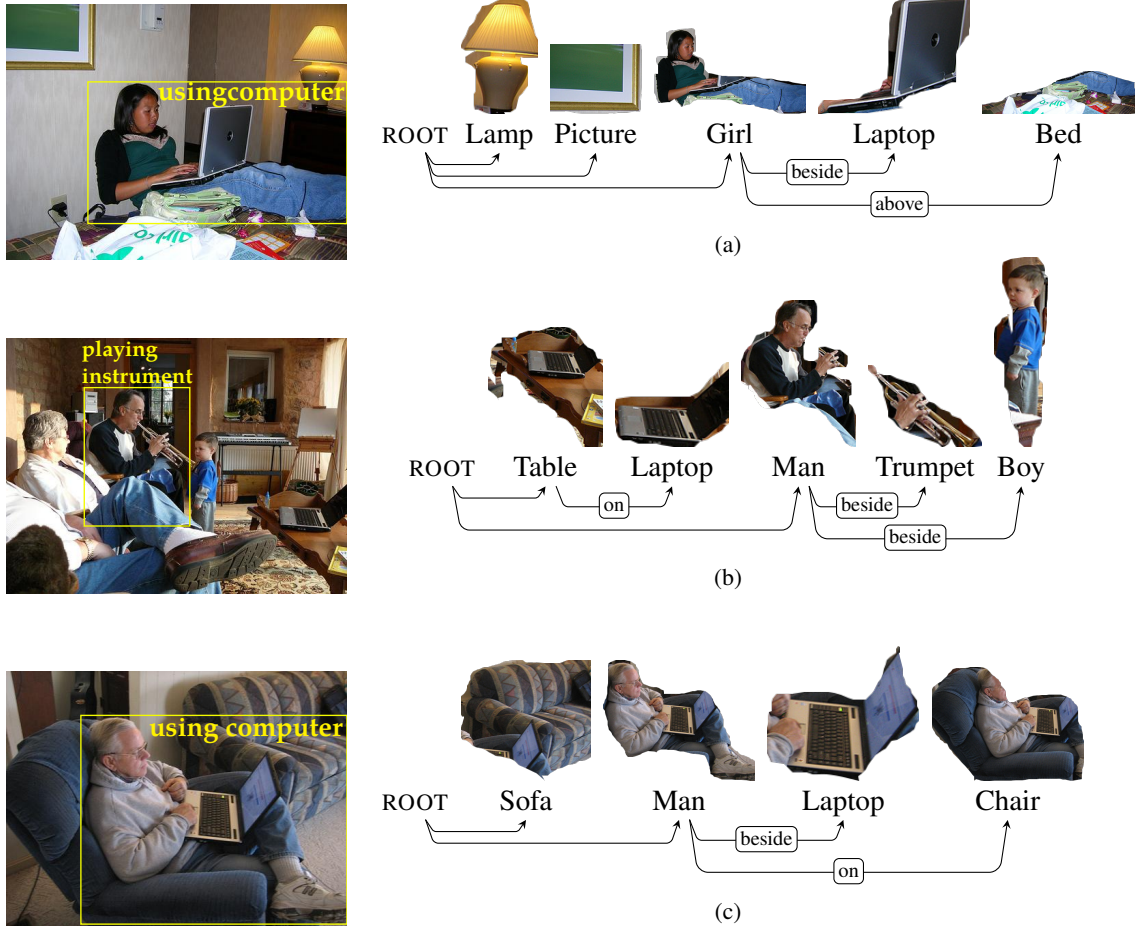
Figure 1: Three examples of images depicting a person and a computer, alongside a respective Visual Dependency Representation for each image. The bag-of-terms representation can be observed in the annotated regions of the Visual Dependency Representations. In (a) and (c) there is a person using a laptop, whereas in (b) the man is actually using the trumpet. The gold-standard action annotation is shown in the yellow bounding box.

and Keller (2013) showed that encoding the spatial relationships between objects in the Visual Dependency Representation helped to generate significantly better descriptions than approaches based on the spatial proximity of objects (Farhadi et al., 2010) or corpus-based models (Yang et al., 2011). In this paper we study whether the Visual Dependency Representation of images can improve the performance of query-by-example image retrieval models. The main finding is that encoding images using the Visual Dependency Representation leads to significantly better retrieval accuracy compared to a bag-of-terms baseline, and that the improvements are most pronounced for transitive verbs.

2 Related Work

2.1 Representing Images

A central problem in image retrieval is how to abstractly represent images (Datta et al., 2008). A bag-of-terms representation of an image is created by grouping visual features, such as color, shape (Shi and Malik, 2000), texture, and interest points (Lowe, 1999), in a vector or as a probability distribution over the features. Image retrieval can then be performed by trying to find the best matchings of terms across an image collection. Spatial Pyramid Matching is an approach to constructing low-level image representations that capture the relationships between features at differently sized partitions of the image (Lazebnik et al., 2006). This approach has proven successful for scene categorisation tasks. An alternative approach to representing images is to learn a mapping (Duygulu et al., 2002; Lavrenko et al.,

110

2003; Guillaumin and Mensink, 2009) between the bags-of-terms and object tags. An image can then be represented as a bag-of-terms and image retrieval is similar to text retrieval (Wu et al., 2012).

In this work, we represent an image as a directed acyclic graph over a set of labeled object region annotations. This representation captures the important spatial relationships between the image regions and makes it possible to distinguish between co-occurring regions and interacting regions.

2.2   Still-Image Action Recognition

One approach to recognizing actions is to learn appearance models for *visual phrases* and use these models to predict actions (Sadeghi and Farhadi, 2011). A visual phrase is defined as the people and the objects they interact with in an action. In this approach, a fixed number of visual phrase models are trained using the deformable parts object detector (Felzenszwalb et al., 2010) and used to perform action recognition.

An alternative approach is to model the relationships between objects in an image, and hence the visible actions, as a Conditional Random Field (CRF), where each node in the field is an object and the factors between nodes correspond to features that capture the relationships between the objects (Zitnick et al., 2013). The factors between object nodes in the CRF include object occurrence, absolute position, person attributes, and the relative location of pairs of objects. This model has been used to generate novel images of people performing actions and to retrieve images of people performing actions.

Most recently, actions have been predicted in images by selecting the most likely verb and object pair given a set of candidate objects detected in an image (Le et al., 2013a). The verb and object is selected amongst those that maximize the distributional similarity of the pair in a large and diverse collection of documents. This approach is most similar to ours but it relies on an external corpus and, depending on the text collections used to train the distributional model, will compound the problem of co-occurrence of objects instead of the relationships between the objects.

The work presented in this paper uses ground-truth annotation for region labels, an assumption similar to (Zitnick et al., 2013), but requires no external data to make predictions of the relationships between objects, unlike the approach of (Le et al., 2013a). The directed acyclic graph representation we propose for images can be seen as a latent representation of the depicted action in the image, where the spatial relationships between the regions capture the different types of actions.

3   Task and Baseline

In this paper we study the task of query-by-example image retrieval within the restricted domain of images depicting actions. More specifically, given an image that depicts a given action, such as *using a computer*, the aim of the retrieval model is to find all other images in the image collection that depict the same action. We define an action as an event involving one or more entities in an image, e.g., *a woman running* or *boy using a computer*, and assume all images have been manually annotated for objects. This assumption means we can explore the utility of the Visual Dependency Representation without the noise introduced by automatic computer vision methods. The data available to the retrieval models can be seen in Figure 1, and Section 5 provides further details about the different sources of data The action label - which is only used for evaluation - is shown in the labelled bounding box, and the Visual Dependency Representation - not used by the baseline model - is shown as a tree at the bottom of the figure.

The main hypothesis explored in this paper is that the accuracy of an image retrieval model will increase if the representation encodes information about the relationships between the objects in images. This hypothesis is tested by encoding images as either an unstructured bag-of-terms representation or as the structured Visual Dependency Representation. The Bag-of-Terms baseline represents the query image and the image collection as an unstructured bags-of-terms vector. All of the models used to test the main hypothesis use the cosine similarity function is to determine the similarity of the query image to other images in the collection, and thus to generate a ranked list from the similarity values.

4 Visual Dependency Representation

The Visual Dependency Representation (VDR) is a structured representation of an image that captures the spatial relationships between pairs of image regions in a directed labelled graph. The Visual Dependency Grammar defines eight possible spatial relationships between pairs of regions, as shown in Table 1. The relationships in the grammar were designed to provide *sufficient* coverage of the types of spatial relationships required to describe the data, and are mathematically defined in terms of pixel overlap, distance between regions, and the angle between regions. The frame of reference for annotating spatial relationships is the image itself and not the object in the image, and angles and distance measurements are taken or estimated from the centroids of the regions. The VDR of an image is created by a trained human annotator in a two-stage process:

1. The annotator draws and labels boundaries around the parts of the image they think contribute to defining the action depicted in the image, and the context within which the action occurs;

2. The annotator draws labelled directed edges between the annotated regions that captures how the relationships between the image convey the action. In Section 4.1, we will explain how to automate the second stage of the process from a collection of labelled region annotations.

In addition to the annotated image regions, a VDR also contains a ROOT node, which acts as a placeholder for the image. In the remainder of this section we describe how a gold-standard VDR is created by a human annotator. The starting point for the VDR in Figure 1(a) is the following set of regions and the ROOT node:

First, the regions are attached to each other based on how the relationship between the objects contributes to the depicted action. In Figure 1(a), the Girl is *using* the Laptop, therefore a labelled directed edge is created from the Girl region to the Laptop region. The spatial relationship is labelled as BESIDE.
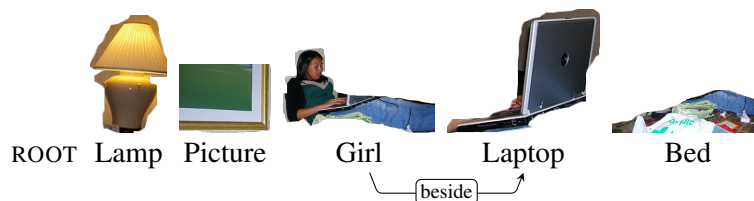
The Girl is also attached to the Bed because the bed supports her body. The spatial relation label is ABOVE because it expresses the spatial relationship between the regions, not the semantic relationship ON. ROOT is attached to the Girl without an edge label to symbolize that she is an actor in the image.
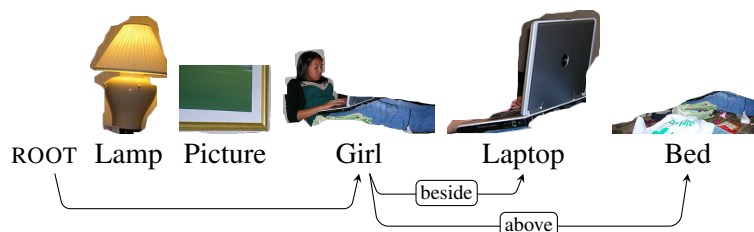
Now the regions that are not concerned with the depicted action are first attached to each other if there is a clear spatial relationship between them (for an example, see Figure 1(b), where the laptop is attached to the table because it is sitting on the table), and then to the ROOT node to signify that they do not play a part in the depicted action. In this example, neither the Lamp nor the Picture are related to the action of using the computer, so they are attached to the ROOT node.
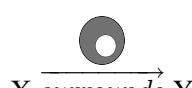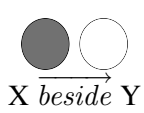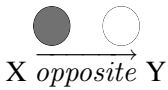
112
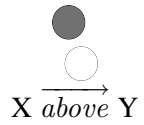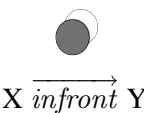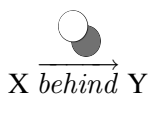
X $\overrightarrow{on}$ Y	More than 50% of the pixels of region X overlap with region Y.	X $\overrightarrow{surrounds}$ Y	The entirety of region X overlaps with region Y.
X $\overrightarrow{beside}$ Y	The angle between the centroid of X and the centroid of Y lies between 315° and 45° or 135° and 225°.	X $\overrightarrow{opposite}$ Y	Similar to *beside*, but used when there X and Y are at opposite sides of the image.
X $\overrightarrow{above}$ Y	The angle between X and Y lies between 225° and 315°.	X $\overrightarrow{below}$ Y	The angle between X and Y lies between 45° and 135°.
X $\overrightarrow{infront}$ Y	The Z-plane relationship between the regions is dominant.	X $\overrightarrow{behind}$ Y	Identical to *infront* except X is behind Y in the Z-plane.

Table 1: Visual Dependency Grammar defines eight relations between pairs of annotated regions. To simplify explanation, all regions are circles, where X is the grey region and Y is the white region. All relations are considered with respect to the centroid of a region and the angle between those centroids.
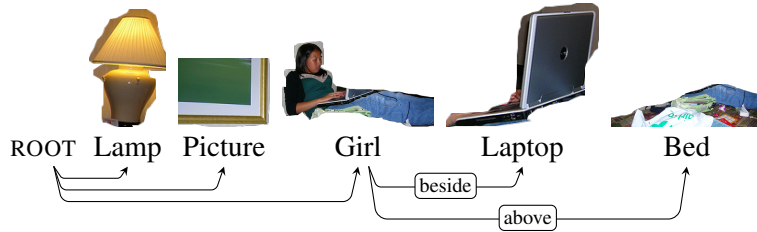
This now forms a completed VDR for the image in Figure 1(a). This structured representation of an image captures the prominent relationship between the girl, the laptop, and the bed. There is no prominent relationship defined between the girl and either the lamp of the picture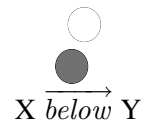, in effect these regions have been relegated to background objects. The central hypothesis underpinning the Visual Dependency Representation is that images that contain similar VDR substructures are more likely to depict the same action than images that only contain the same set of objects. For example, the VDR for Figure 1(a) correctly captures the relationship between the people and the laptops, whereas this relationship is not present in Figure 1(b), where the person is playing a trumpet.

4.1 Predicting Visual Dependency Representations

We follow the approach of Elliott and Keller (2013) and predict the VDR y of an image over a collection of labelled region annotations $\mathbf{x}$. This task is framed as a supervised learning problem, where the aim is to construct a Maximum Spanning Tree from a fully-connected directed weighted graph over the labelled regions (McDonald et al., 2005). Reducing the fully-connected graph to the Maximum Spanning Tree removes the region–region edges that are not important in defining the prominent relationships between the regions in an image. The score of the VDR y over the image regions is calculated as the sum of the scores of the directed labelled edges:

$$score(\mathbf{x}, y) = \sum_{(a,b) \in y} \mathbf{w} \cdot \mathbf{f}(a, b) \qquad (1)$$

where the score of an edge between image regions a and b is calculated using a vector of weighted feature functions $\mathbf{f}$. The feature functions characterize the image regions and the edge between pairs of regions, and include: the labels of the regions and the spatial relation annotated on the edge; the (normalized) distance between the centroids of the regions; the angle formed between the annotated regions, which is

mapped onto the set of spatial relations; the relative size of the region compared to the image; and the distance of the region centroid from the center of the image.

The model is trained over i instances of region-annotated images $\mathbf{x_i}$ associated with human-created VDR structures y_i, $I_{train} = \{\mathbf{x_i}, y_i\}$. The score of each edge a, b is calculated by applying the feature functions to the data associated with that edge, and this is performed over each edge in a VDR to obtain a score for a complete gold-standard structure. The parameters of the weight vector w are iteratively adjusted to maximise the score of the gold-standard structures in the training data using the Margin Infused Relaxation Algorithm (Crammer and Singer, 2002).

The test data contains i instances of region-annotated images with image regions $\mathbf{x_i}$, $I_{test} = \{\mathbf{x_i}\}$. The parsing model computes the highest scoring structure $\hat{y}_i$ for each instance in the test data by scoring each possible directed edge between pairs of regions in $\mathbf{x_i}$. This process forms a fully-connected graph over the image regions, from which the Maximum Spanning Tree is taken and returned as the predicted VDR.

We evaluate the performance of this VDR prediction model by comparing how well it can recover the manually created trees in the data set. This evaluation is performed on the development data in a 10-fold cross validation setting where each fold of the data is split 80%/10%/10%. Unlabelled directed accuracy means the model correctly proposes an edge between a pair of regions in the correct direction; Labelled directed accuracy means it additionally proposes the correct edge label. The baseline approach is to assume no latent image structure and attach all image regions to the ROOT node of the VDR; this achieves 51.6% labelled and unlabelled directed attachment accuracy. The accuracy of our automatic approach to VDR prediction is 61.3% labelled and 68.8% unlabelled attachment accuracy.

4.2 Comparing Visual Dependency Representations

It remains to define how to compare the Visual Dependency Representation of a pair of images. The most obvious approach is to use the labelled directed accuracy measurement used for the VDR prediction evaluation in the previous section, but we did not find significant improvements in retrieval accuracy using this method. We hypothesise that the lack of weight given to the edges between nodes in the Visual Dependency Representation results in this comparison function not distinguishing between object–object relationships that matter, such as PERSON $\overrightarrow{beside}$ BIKE, compared to ROOT $\rightarrow$ TREES. The former is a potential person–object relationship that explains the depicted event, whereas the latter is only a background object.

The approach we adopted in this paper is to compare Visual Dependency Representations of images by decomposing the structure into a set of labelled and a unlabelled parent–child subtrees in a depth-first traversal of the VDR. The decomposition process allows use to use the same similarity function as the Bag-of-Terms baseline model, removing the confound of choosing different similarity functions. The subtrees can be transformed into tokens and these tokens can be used as weighted terms in a vector representation. An example of a labelled transformation is shown below:

$$\boxed{above}$$
Girl Bed $\rightarrow$ Girl_above_Bed
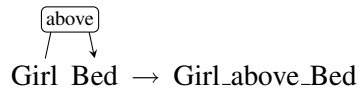
We now demonstrate the outcome of comparing images represented using either a vector that concatenates the decomposed transformed VDR and bag-of-terms, or a vector that contains only the bag-of-terms. In this demonstration, each term has a *tf-idf* weight of 1. The first illustration (*Similar*) compares images that depict the same underlying action: Figure 1 (a) and (c). The second illustration (*Dissimilar*) compares images that depict different actions: Figure 1 (a) and (b).

$$Similar : \cos(\text{VDR}_a, \text{VDR}_c) = 0.56 > \cos(\text{Bag}_a, \text{Bag}_c) = 0.52$$
$$Dissimilar : \cos(\text{VDR}_b, \text{VDR}_a) = 0.201 \ll \cos(\text{Bag}_b, \text{Bag}_a) = 0.4$$

It can be seen that when the images represent the same action, the decomposed VDR increases the similarity of the pair of images compared to the bag-of-terms representation; and when images do not

represent the same action, the decomposed VDR yields a lower similarity than the bag-of-terms representation. These illustrations confirm that Visual Dependency Representations can be used to distinguish the difference between presence or absence of objects, and the prominent relationships between objects.

5 Data

We use an existing dataset of VDR-annotated images to study whether modelling the structure of an image can improve image retrieval in the domain of action depictions. The data set of Elliott and Keller (2013) contains 341 images annotated with region annotations, three visual dependency representations per image (making a total of 1,023 instances), and a ground-truth action label for each image. An example of the annotations can be seen in Figure 1. The image collection is drawn from the PASCAL Visual Object Classification Challenge 2011 action recognition taster and covers a set of 10 actions (Everingham et al., 2011): riding a bike, riding a horse, reading, running, jumping, walking, playing an instrument, using a computer, taking a photo, and talking on the phone.

Image Descriptions

Each image is associated with three human-written descriptions collected from untrained annotators on Amazon Mechanical Turk. The descriptions do not form any part of the models presented in the current paper; they were used in the automatic image description task of Elliott and Keller (2013). Each description contains two sentences: the first sentence describes the action depicted in the image, and the second sentence describes other objects not involved in the action. A two sentence description of an image helps distinguish objects that are central to depicting the action from objects that may be distractors.

Region Annotations

The images contain human-drawn labelled region annotations. The annotations were drawn using the LabelMe toolkit, which allows for arbitrary labelled polygons to be created over an image (Russell et al., 2008). The annotated regions were restricted to those present in at least one of three human-written descriptions. To reduce the effects of label sparsity, frequently occurring equivalent labels were conflated, i.e., man, child, and boy $\rightarrow$ person; bike, bicycle, motorbike $\rightarrow$ bike; this reduced the object label vocabulary from 496 labels to 362 labels. The data set contains a total of 5,034 region annotations, with a mean of 4.19 ± 1.94 annotations per image.

Visual Dependency Representations

Recall that each image is associated with three descriptions, and that people were free to decide how to describe the action and background of the image. The differences between how people describe images leads to the creation of one Visual Dependency Representation per image–description pair in the data set, resulting in a total of 1,023 instances. The process for creating a visual dependency representation of an image is described in Section 4. The annotated dataset comprises a total of 5,748 spatial relations, corresponding to a mean of 4.79 ± 3.51 relations per image. Elliott and Keller (2013) report inter-annotator agreement on a subset of the data at 84% agreement for labelled directed attachments and 95.1% for unlabelled directed attachments.

Action Labels

The original PASCAL action recognition dataset contains ground truth action class annotations for each image. These annotations are in the form of labelled bounding boxes around the person performing the action in the image. The action labels are only used as the gold-standard relevance judgements for the query-by-example image retrieval experiments.

6 Experiments

In this section we present the results of a query-by-example image retrieval experiment to determine the utility of the Visual Dependency Representation compared to a bag-of-terms representation. In this
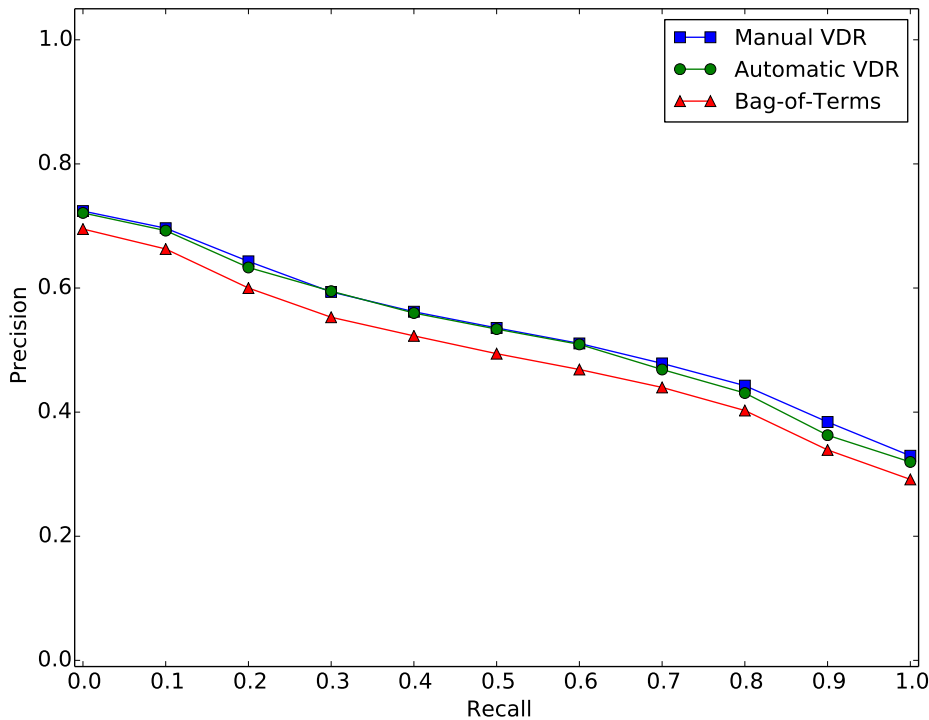
Figure 2: Average 11-point precision/recall curves show that the VDR-based retrieval models are consistently better than the Bag-of-Terms model.

experiment, a single image (the query image) is used to rank the images in the test collection, where the goal is to construct a ranking where the top images depict the same action as the query image.

6.1 Protocol

The image retrieval experiment is performed using 10-fold cross-validation in the following manner. The 341 images in the dataset are randomly partitioned into 80%/10%/10% splits, resulting in 1011 test queries[1]. For each query we compute average precision and Precision@10 of the ranked list, and use the resulting values to test the statistical significance of the results.

The *training set* is used to train the VDR prediction model and to estimate inverse document frequency statistics. During the training phase, the VDR-based models have access to region boundaries, region labels and three manually-created VDRs for each training image. In the *test set*, all models have access to the region boundaries and labels for each image. Each image in the test set forms a query and the models produce a ranked list of the remaining images in the test collection. Images are marked for relevance as follows: a image at rank r is considered *relevant* if it has the same action label as the query image; otherwise it is *non-relevant*. The *dev set* was used to experiment with different matching functions and to optimise the feature functions used in the VDR prediction model.

6.2 Models

We compare the retrieval accuracy of three approaches: Bag-of-Terms uses an unstructured representation for each image. A *tf-idf* weight is assigned to each region label in an image, and the cosine measure is used to calculate the similarity of images. This model allows us to compare the usefulness of a structured vs. unstructured image representation. Automatic VDR is a model using the VDR prediction method from Section 4.1, and Manual VDR uses the gold-standard data described in Section 5. Both

---

[1]Recall there are three Visual Dependency Representations for each image. The partitions are the same as those used in the VDR prediction experiment in Section 4.1

	MAP	P@10
Manual VDR	$0.514^{\star\dagger}$	$0.454^\star$
Automatic VDR	$0.508^\star$	$0.451^\star$
Bag-of-Terms	0.467	0.415

Table 2: Overall Mean Average Precision and Precision@10 images. The VDR-based models are significantly better than the Bag-of-Terms model, supporting the hypothesis that modelling the structure of an image using the Visual Dependency Representation is useful for image retrieval. $\star$: significantly different than Bag-of-Terms at $p < 0.01$; $\dagger$: significantly different than Automatic VDR at $p < 0.01$.

of the VDR-based models have a tf-idf weight assigned to the transformed decomposed terms and the cosine similarity measure is used to calculate the similarity of images.

6.3 Results

Figure 2(a) shows the interpolated precision/recall curve and Table 2 shows the Mean Average Precision (MAP) and Precision at 10 retrieved images (P@10). The MAP of the Automatic VDR model increases by 8.8% relative to the Bag-of-Terms model, and a relative improvement up to 10.1% would possible if we had a better structure prediction model, as evidenced by Manual VDR. Furthermore, if we assume a user will only view the top results returned by the retrieval model, then P@10 increases by 8.6% when we model the structure of an image, relative to using an unstructured representation; a relative improvement of up to 9.4% would be possible if we had a better image parser.

To determine whether the differences are statistically significant, we perform the Wilcoxon Signed Ranks Test on the average precision and P@10 values over the 1011 queries in our cross-validation data set. The results support the main hypothesis of this paper: structured image representations allow us to find images depicting actions more accurately than the standard bag-of-terms representation. We find significant differences in average precision and P@10 between the Bag-of-Terms baseline and both Automatic VDR ($p < 0.01$) and Manual VDR ($p < 0.01$). This suggests that structure is very useful in the query-by-example scenario. We find a significant difference in average precision between Automatic VDR and Manual VDR ($p < 0.01$), but no difference in P@10 between Automatic VDR and Manual VDR ($p = 0.442$).

6.4 Retrieval Performance by Type of Action and Verb

We now analyse whether image structure is useful when the action does not require a direct object. The analysis presented here compares the Bag-of-Terms model against the Automatic VDR model because there was no significant difference in P@10 between the Automatic and Manual VDR models. Table 3 shows the MAP and Precision@10 per type of action. Figure 3 shows the precision/recall curves for (a) transitive verbs, (b) intransitive verbs, and (c) light verbs.

In Figure 3(a), it can be seen that the actions that can be classified as transitive verbs benefit from exploiting the structure encoded in the Visual Dependency Representation. The only exception is for the action *to read*, which frequently behaves as an intransitive verb: *the man reads on a train*. The consistent improvement in both the entirety of the ranked list and at the top of the ranked list can be seen in the MAP and P@10 results in Table 3.

Figure 3(b) shows that there is a small increase in retrieval performance for intransitive verbs compared to the transitive verbs. We conjecture this is because there are fewer objects to annotate in an image when the verb does not require a direct object. The summary results for the intransitive verbs in Table 3 confirm the small but insignificant increase in MAP and P@10.

Finally, the light verbs, shown in Figure 3(c), exhibit variable behaviour in retrieval performance. One reason for this could be that if the light verb encodes information about the object, as in *using a computer*, then the computer can be annotated in the image, and thus it acts as a transitive verb. Conversely, when
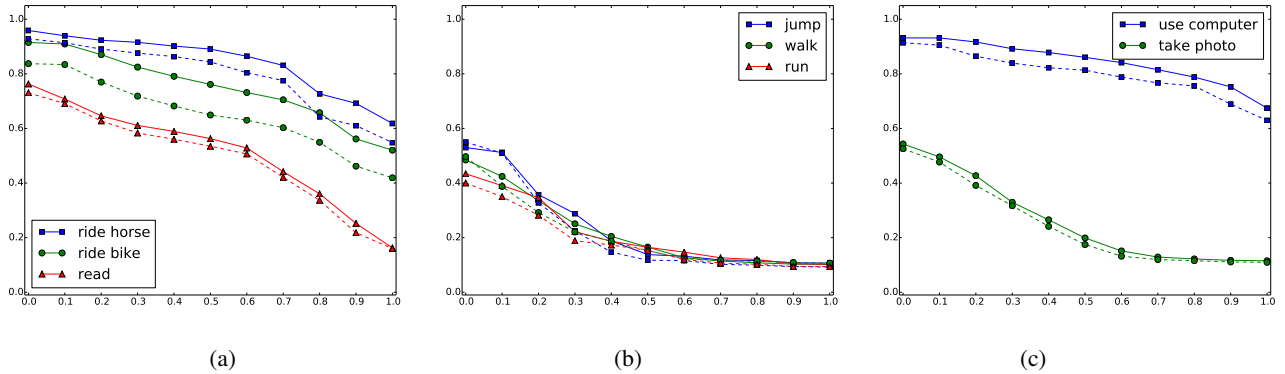
Figure 3: Precision/recall curves grouped by the type of verb. The solid lines represent the Automatic VDR model; the dashed lines represent the Bag-of-Terms model; y-axis is Precision, and the x-axis is Recall. (a) Images depicting transitive verbs benefit the most from the Visual Dependency Representation and are easiest to retrieve. (b) Intransitive verbs are difficult to retrieve and there is is a negligible improvement in performance when using Visual Dependency Representation. (c) Light verbs benefit from the Visual Dependency Representation depending on the type of the object involved in the action.

	MAP		P@10	
	VDR	Bag	VDR	Bag
Ride bike	0.721*	0.601	0.596*	0.513
Ride horse	0.833*	0.768	0.787*	0.726
Talk on phone	0.762*	0.679	0.666*	0.582
Play instrument	0.774*	0.705	0.634*	0.586
Read	0.483	0.454	0.498	0.475
Walk	0.198	0.186	0.184	0.174
Run	0.193	0.165	0.151	0.132
Jump	0.211	0.189	0.142	0.136
Use computer	0.814*	0.761	0.694*	0.648
Take photo	0.241	0.223	0.212	0.198

Table 3: Mean Average Precision and Precision@10 for each action in the data set, grouped into transitive (top), intransitive (middle), and light (bottom) verbs. VDR is the Automatic VDR model and Bag is the Bag-of-Terms model. It can be seen that the Automatic VDR retrieval model is consistently better than the Bag-of-Terms model on both MAP and Precision@10. *: the Automatic VDR model is significantly different than Bag-of-Terms at $p < 0.01$.

the light verb conveys information about the outcome of the event, as in the action *take a photograph*, the outcome is rarely possible to annotate in an image, and so no improvements can be gained from structured image representations.

6.5 Discussion

In our experiments we observed that all models can achieve high precision at very low levels of recall. We found that this happens for testing images that are almost identical to the query image. For such images, objects that are unrelated to the target action form an effective context, which allows this image to be placed at the top of the ranking. However, near-identical images are relatively rare, and performance degrades for higher levels of recall.

It is surprising that image retrieval using automatically predicted VDR model is statistically indistinguishable from the manually crafted VDR model, given the relatively low accuracy of our VDR prediction model: 61.3% by the labelled dependency attachment accuracy measure. One possible explanation could be that not all parts of the VDR structure are useful for retrieval purposes, and our VDR prediction model does well on the useful ones. This observation also suggests that we are unlikely to achieve better retrieval performance by continuing to improve the accuracy of VDR prediction. We believe a more promising direction is refining the current formulation of the VDR, and exploring more sophisticated ways to measure the similarity of two structured representations.

7 Conclusion

In this paper we argued that a limiting factor of retrieving images depicting actions is the unstructured bag-of-terms representation typically used for images. In a bag-of-terms representation, images that share similar sets of regions are deemed to be related even when the depicted actions are different. We proposed that representing an image using the Visual Dependency Representation (VDR) can prevent this type of misclassification in image retrieval. The VDR of an image captures the region–region relationships that explain what is happening in an image, and it can be automatically predicted from a region-annotated image.

In a query-by-example image retrieval task, we found that representing images as automatically predicted VDRs resulted in statistically significant 8.8% relative improvement in MAP and 8.6% relative improvement in Precision@10 compared to a Bag-of-Terms model. There was a significant difference in MAP when using manually or automatically predicted image structures, but no difference in the Precision@10, suggesting that the proposed automatic prediction model is accurate enough for retrieval purposes. Future work will focus on using automatically generated visual input, such as the output of the image tagger (Guillaumin and Mensink, 2009), or an automatic object detector (Felzenszwalb et al., 2010), which will make it possible to tackle image ranking tasks (Hodosh et al., 2013). It would also be interesting to explore alternative structure prediction methods, such as predicting the relationships using a conditional random field (Zitnick et al., 2013), or by leveraging distributional lexical semantics (Le et al., 2013b).

Acknowledgments

References

Moshe Bar and Shimon Ullman. 1996. Spatial Context in Recognition. *Perception*, 25(3):343–52, January.

I Biederman. 1972. Perceiving real-world scenes. *Science*, 177(4043):77–80.

Koby Crammer and Yoram Singer. 2002. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research*, 2:265–292.

Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z. Wang. 2008. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys*, 40(2):1–60.

P Duygulu, Kobus Barnard, J F G de Freitas, and David A Forsyth. 2002. Object Recognition as Machine Translation: Learning a Lexicon for a Fixed Image Vocabulary. In *Proceedings of the 7th European Conference on Computer Vision*, pages 97–112, Copenhagen, Denmark.

Desmond Elliott and Frank Keller. 2013. Image Description using Visual Dependency Representations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1292–1302, Seattle, Washington, U.S.A.

Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. 2011. The PASCAL Visual Object Classes Challenge 2011.

Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. 2010. Every picture tells a story: generating sentences from images. In *Proceedings of the 15th European Conference on Computer Vision*, pages 15–29, Heraklion, Crete, Greece.

P F Felzenszwalb, R B Girshick, D McAllester, and D Ramanan. 2010. Object Detection with Discriminatively Trained Part-Based Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645.

Matthieu Guillaumin and Thomas Mensink. 2009. Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In *IEEE 12th International Conference on Computer Vision*, pages 309–316, Kyoto, Japan.

Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics. *Journal of Artificial Intelligence Research*, 47:853–899.

Victor Lavrenko, R Manmatha, and Jiwoon Jeon. 2003. A Model for Learning the Semantics of Pictures. In *Advances in Neural Information Processing Systems 16*, Vancouver and Whistler, British Columbia, Canada.

S. Lazebnik, C. Schmid, and J. Ponce. 2006. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2169–2178, New York, NY, USA.

DT Le, R Bernardi, and Jasper Uijlings. 2013a. Exploiting language models to recognize unseen actions. In *Proceedings of the International Conference on Multimedia Retrieval*, pages 231–238, Dallas, Texas, U.S.A.

DT Le, Jasper Uijlings, and Raffaella Bernardi. 2013b. Exploiting language models for visual recognition. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 769–779, Seattle, Washington, U.S.A.

D G Lowe. 1999. Object recognition from local scale-invariant features. In *Proceedings of the International Conference on Computer Vision*, pages 1150–1157, Washington, D.C., USA.

Ryan McDonald, Fernando Pereira, Kiril Ribarov, and Jan Hajič. 2005. Non-projective dependency parsing using spanning tree algorithms. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 523–530, Vancouver, British Columbia, Canada.

Bryan C. Russell, Antonio Torralba, Kevin P. Murphy, and William T. Freeman. 2008. LabelMe: A Database and Web-Based Tool for Image Annotation. *International Journal of Computer Vision*, 77(1-3):157–173.

Mohammad A Sadeghi and Ali Farhadi. 2011. Recognition Using Visual Phrases. In *2011 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1745–1752, Colorado Springs, Colorado, U.S.A.

Jianbo Shi and Jitendra Malik. 2000. Normalized Cuts and Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, August.

Lei Wu, Rong Jin, and Anil K Jain. 2012. Tag Completion for Image Retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(3):716–727.

Yezhou Yang, Ching Lik Teo, Hal Daumé III, and Yiannis Aloimonos. 2011. Corpus-Guided Sentence Generation of Natural Images. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 444–454, Edinburgh, Scotland, UK.

CL Zitnick, Devi Parikh, and Lucy Vanderwende. 2013. Learning the Visual Interpretation of Sentences. In *IEEE International Conference on Computer Vision*, pages 1681–1688, Sydney, Australia.