Image Description using Visual Dependency Representations

Desmond Elliott
School of Informatics
University of Edinburgh
d.elliott@ed.ac.uk

Frank Keller
School of Informatics
University of Edinburgh
keller@inf.ed.ac.uk

Abstract

Describing the main event of an image involves identifying the objects depicted and predicting the relationships between them. Previous approaches have represented images as unstructured bags of regions, which makes it difficult to accurately predict meaningful relationships between regions. In this paper, we introduce visual dependency representations to capture the relationships between the objects in an image, and hypothesize that this representation can improve image description. We test this hypothesis using a new data set of region-annotated images, associated with visual dependency representations and gold-standard descriptions. We describe two template-based description generation models that operate over visual dependency representations. In an image description task, we find that these models outperform approaches that rely on object proximity or corpus information to generate descriptions on both automatic measures and on human judgements.

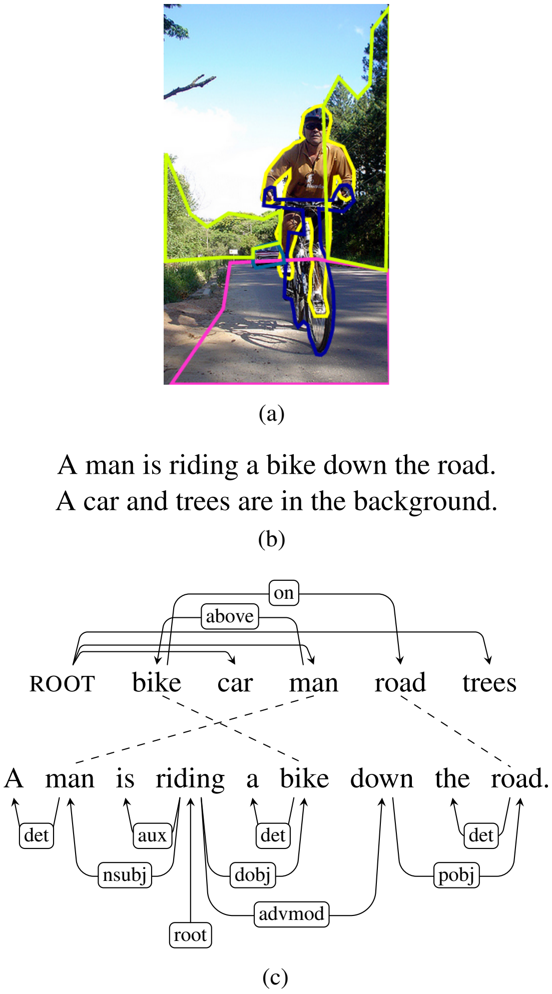
1 Introduction

Humans are readily able to produce a description of an image that correctly identifies the objects and actions depicted. Automating this process is useful for applications such as image retrieval, where users can go beyond keyword-search to describe their information needs, caption generation for improving the accessibility of existing image collections, story illustration, and in assistive technology for blind and partially sighted people. Automatic image description presents challenges on a number of levels: recognizing the objects in an image and their attributes are difficult computer vision problems; while determining how the objects interact, which relationships hold between them, and which events are depicted requires considerable background knowledge.

Previous approaches to automatic description generation have typically tackled the problem using an object recognition system in conjunction with a natural language generation component based on language models or templates (Kulkarni et al., 2011; Li et al., 2011). Some approaches have utilised the visual attributes of objects (Farhadi et al., 2010), generated descriptions by retrieving the descriptions of similar images (Ordonez et al., 2011; Kuznetsova et al., 2012), relied on an external corpus to predict the relationships between objects (Yang et al., 2011), or combined sentence fragments using a tree-substitution grammar (Mitchell et al., 2012).

A common aspect of existing work is that an image is represented as a bag of image regions. Bags of regions encode which objects co-occur in an image, but they are unable to express how the regions relate to each other, which makes it hard to describe what is happening. As an example, consider Figure 1a, which depicts a man riding a bike. If the man was instead repairing the bike, then the bag-of-regions representation would be the same, even though the image would depict a different action and would have to be described differently. This type of co-occurrence of regions indicates the need for a more structured image representation; an image description system that has access to structured repre-

Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 1292–1302,
Seattle, Washington, USA, 18-21 October 2013. ©2013 Association for Computational Linguistics



(a)

A man is riding a bike down the road.
A car and trees are in the background.

(b)

(c)

Figure 1: (a) Image with regions marked up: BIKE, CAR, MAN, ROAD, TREES; (b) human-generated image description; (c) visual dependency representation expressing the relationships between MAN, BIKE, and ROAD aligned to the syntactic dependency parse of the first sentence in the human-generated description (b).

sentations would be able to correctly infer the action that is taking place, such as the distinction between repairing or riding a bike, which would greatly improve the descriptions it is able to generate.

In this paper, we introduce *visual dependency representations* (VDRs) to represent the structure of images. This representation encodes the geometric relations between the regions of an image. An example can be found in Figure 1c, which depicts the VDR for Figure 1a. It encodes that the MAN is above the BIKE, and that the BIKE is on the ROAD. These relationships make it possible to infer that the man is riding a bike down the road, which corresponds to the first sentence of the human-generated image description in Figure 1b.

In order to test the hypothesis that structured image representations are useful for description generation, we present a series of template-based image description models. Two of these models are based on approaches in the literature that represent images as bags of regions. The other two models use visual dependency representations, either on their own or in conjunction with gold-standard image descriptions at training time.

We find that descriptions generated using the VDR-based models are significantly better than those generated using bag-of-region models in automatic evaluations using smoothed BLEU scores and in human judgements. The BLEU score improvements are found at bi-, tri-, and four-gram levels, and humans rate VDR-based image descriptions 1.2 points above the next-best model on a 1–5 scale.

Finally, we also show that the benefit of the visual dependency representation is maintained when image descriptions are generated from automatically parsed VDRs. We use a modified version of the edge-factored parser of McDonald et al. (2005) to predict VDRs over a set of annotated object regions. This result reaffirms the potential utility of this representation as a means to describe events in images. Note that throughout the paper, we work with gold-standard region annotations; this makes it possible to explore the effect of structured image representations independently of automatic object detection.

2 Visual Dependency Representation

In analogy to dependency grammar for natural language syntax, we define *Visual Dependency Grammar* to describe the spatial relations between pairs of image regions. A directed arc between two regions is labelled with the spatial relationship between those regions, defined in terms of three geometric properties: pixel overlap, the angle between regions, and the distance between regions. Table 1 presents a detailed explanation of the spatial relationships defined in the grammar.

A visual dependency representation of an image is constructed by creating a directed acyclic graph

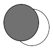
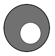
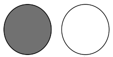
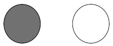
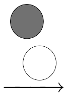
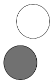
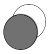
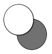
 $X \overrightarrow{\textit{on}} Y$	More than 50% of the pixels of region X overlap with region Y.[1]
 $X \overrightarrow{\textit{surrounds}} Y$	The entirety of region X overlaps with region Y.
 $X \overrightarrow{\textit{beside}} Y$	The angle between the centroid of X and the centroid of Y lies between 315° and 45° or 135° and 225°.
 $X \overrightarrow{\textit{opposite}} Y$	Similar to *beside*, but used when there X and Y are at opposite sides of the image.
 $X \overrightarrow{\textit{above}} Y$	The angle between X and Y lies between 225° and 315°.
 $X \overrightarrow{\textit{below}} Y$	The angle between X and Y lies between 45° and 135°.
 $X \overrightarrow{\textit{infront}} Y$	The Z-plane relationship between the regions is dominant.
 $X \overrightarrow{\textit{behind}} Y$	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. We follow the definition of the unit circle, in which 0° lies to the right and a turn around the circle is counter-clockwise.

over the set of regions in an image using the spatial relationships in the Visual Dependency Grammar. It is created from a region-annotated image and a corresponding image description by first identifying the central actor of the image. The central actor is the person or object carrying out the depicted action; this typically corresponds to the subject of the sentence describing the image. The region corresponding to the central actor is attached to the ROOT node of the graph. The remaining regions are then attached based on their relationship with either the actor or the other regions in the image as they are mentioned in the description. Each arc introduced is labelled with one of the spatial relations defined in the grammar, or with no label if the region is not described in relation to anything else in the image.

As an example of the output of this annotation process, consider Figure 1a, its description in 1b, and its VDR in 1c. Here, the MAN is the central actor in the image, as he is carrying out the depicted action (riding a bike). The region corresponding to MAN is therefore attached to ROOT without a spatial relation. The BIKE region is then attached to the MAN region using the $\overrightarrow{\textit{above}}$ relation and BIKE is attached to the ROAD with the $\overrightarrow{\textit{on}}$ relation. In the second sentence of the description, CAR and TREES are mentioned without a relationship to anything else in the image, so they are attached to the ROOT node. If these regions were attached to other regions, such as CAR $\overrightarrow{\textit{above}}$ ROAD then this would imply structure in the image that is not conveyed in the description.

2.1 Data

Our data set uses the images from the PASCAL Visual Object Classification Challenge 2011 action recognition taster competition (Everingham et al., 2011). This is a closed-domain data set containing images of people performing ten types of actions, such as making a phone call, riding a bike, and taking a photo. We annotated the data set in a three-step process: (1) collect a description for each image; (2) annotate the regions in the image; and (3) create a visual dependency representation of the image. Note that Steps (2) and (3) are dependent on the image description, as both the region labels and the relations between them are derived from the description.

2.2 Image Descriptions

We collected three descriptions of each image in our data set from Amazon Mechanical Turk. Workers were asked to describe an image in two sentences. The first sentence describes the action in the image, the person performing the action and the region involved in the action; the second sentence describes any other regions in the image not directly involved in the action. An example description is given in Figure 1b.

A total of 2,424 images were described by three workers each, resulting in a total of 7,272 image de-

[1] As per the PASCAL VOC definition of overlap in the object detection task (Everingham et al., 2011).

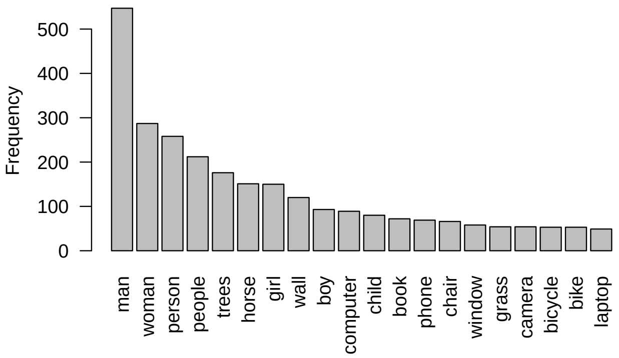


Figure 2: Top 20 annotated regions.

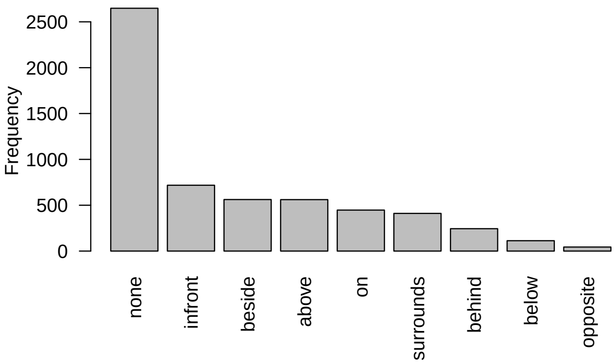


Figure 3: Distribution of the spatial relations.

scriptions. The workers, drawn from those registered in the US with a minimum HIT acceptance rate of 95%, described an average of 145 ± 93 images; they were encouraged to describe fewer than 300 images each to ensure a linguistically diverse data set. They were paid $0.04 per image and it took on average 67 ± 123 seconds to describe a single image. The average length of a description was 19.9 ± 6.5 words in a range of 8–50 words. Dependency parses of the descriptions were produced using the MST-Parser (McDonald et al., 2005) trained on sections 2-21 of the WSJ portion of the Penn Treebank.

2.3 Region Annotations

We trained two annotators to draw polygons around the outlines of the regions in an image using the LabelMe annotation tool (Russell et al., 2008). The regions annotated for a given image were limited to those mentioned in the description paired with the image. Region annotation was performed on a subset of 341 images and resulted in a total of 5,034 annotated regions with a mean of 4.19 ± 1.94 annotations per image. A total of 496 distinct labels were used to label regions. Figure 2 shows the distribution of the top 20 region annotations in the data; people-type regions are the most commonly annotated regions. Given the prevalence of labels referring to the same types of regions, we defined 26 sets of equivalent labels to reduce label sparsity (e.g., BIKE was considered equivalent to BICYCLE). This normalization process reduced the size of the region label vocabulary from 496 labels to 362 labels. Inter-annotator agreement was 74.3% for region annotations, this was measured by computing polygon overlap over the annotated regions.

2.4 Visual Dependency Representations

The same two annotators were trained to construct gold-standard visual dependency representations for annotated image–description pairs. The process for creating a visual dependency representation of an image is described earlier in this section of the paper. The 341 region-annotated images resulted in a set of 1,023 visual dependency representations. The annotated data set comprised a total of 5,748 spatial relations, corresponding to a mean of 4.79 ± 3.51 relations per image. Figure 3 shows the distribution of spatial relation labels in the data set. It can be seen that the majority of regions are attached to the ROOT node, i.e., they have the relation label *none*. Inter-annotator agreement on a subset of the data was measured at 84% agreement for labelled dependency accuracy and 95.1% for unlabelled dependency accuracy. This suggests the task of generating visual dependency representations can be performed reliably by human annotators. We induced an alignment between the annotated region labels and words in the image description using simple lexical matching augmented with WordNet hyponym lookup. See Figure 1c for an example of the alignments.

3 Image Description Models

We present four template-based models for generating image descriptions in this section. Table 2

	Regions	VDR	External Corpus	Parallel text
PROXIMITY	✓			
CORPUS	✓		✓	
STRUCTURE	✓	✓		
PARALLEL	✓	✓		✓

Table 2: The data available to each model at training time.

T_1	DT O_i AUX REL DT O_j. T_5?
T_2	There AUX also $\{\text{DT O}_i\}_{i=1}^{\lvert unrelated \rvert}$ in the image.
T_3	DT O_i AUX REL DT O_j REL DT O_k. T_5?
T_4	REL DT O_j.
T_5	PRP AUX $\{\text{REL DT O}_i\}_{i=1}^{\lvert dependents \rvert}$.

Table 3: The language generation templates.

presents an overview of the amount of information available to each model at training time, ranging from only the annotated regions of an image to using visual dependency representation of an image aligned with the syntactic dependency representation of its description. At test time, all models have access to image regions and their labels, and use these to generate image descriptions. Two of the models also have access to VDRs at test time, allowing us to test the hypothesis that image structure is useful for generating good image descriptions.

The aim of each model is to determine what is happening in the image, which regions are important for describing it, and how these regions relate to each other. Recall that all our images depict actions, and that the gold-standard annotation was performed with this in mind. A good description therefore is one that relates the main actors depicted in the image to each other, typically through a verb; a mere enumeration of the regions in the image is not sufficient. All models attempt to generate a two-sentence description, as per the gold standard descriptions.

In the remainder of this section, we will use Figure 1 as a running example to demonstrate the type of language each model is capable of generating. All models share the set of templates in Table 3.

3.1 PROXIMITY

PROXIMITY is based on the assumption that people describe the relationships between regions that are near each other. It has access to only the annotated image regions and their labels.

Region–region relationships that are potentially relevant for the description are extracted by calculating the proximity of the annotated regions. Here, o_i is the subject region, o_j is the object region, and s_{ij} is the spatial relationship between the regions. Let $\text{R} = \{(\text{o}_i, \text{s}_{ij}, \text{o}_j), \dots\}$ be the set of possible region–region relationships found by calculating the nearest neighbour of each region in Euclidean space between the centroids of the polygons that mark the region boundaries. The tuple with the subject closest to the centre of the image is used to describe what is happening in the image, and the remaining regions are used to describe the background.

The first sentence of the description is realised with template T_1 from Table 3. o_i is the label of the subject region and o_j is the label of the object region. DT is a simple determiner chosen from {the, a}, depending on whether the region label is a plural noun; AUX is either {is, are}, depending on the number of the region label; and REL is a word to describe the relationship between the regions. For this model, REL is the spatial relationship between the centroids chosen from {above, below, beside}, depending on the angle formed between the region centroids, using the definitions in Table 1. The second sentence of the description is realised with template T_2 over the subjects o_i in R that were not used in the first sentence. An example of the language generated is:

(1) The man is beside the bike. There is also a road, a car, and trees in the image.

With the exception of visual attributes to describe size, colour, or texture, this model is based on the approach described by Kulkarni et al. (2011).

3.2 CORPUS

The biggest limitation of PROXIMITY is that regions that are near each other are not always in a relevant relationship for a description. For example, in Figure 1, the BIKE and the CAR regions are nearest neighbours but they are unlikely to be described as being in an relationship by a human annotator. The model CORPUS addresses this issue by using an external text corpus to determine which pairs of regions are likely to be in a describable relationship. Furthermore, CORPUS can generate verbs instead of spatial relations between regions, leading to more human-like descriptions. CORPUS is based on Yang et al. (2011), except we do not use scene type (indoor, outdoor, etc.) as part of the model. At training time, the model has access to the annotated image regions and labels, and to the dependency-parsed version of the English Gigaword Corpus (Napoles et al., 2012). The corpus is used to extract subject–verb–object subtrees, which are then used to predict the best pairs of regions, as well as the verb that relates the regions.

The set of region–region relationships $\mathrm{R} = \{(\mathrm{o}_i, \mathrm{v}_{ij}, \mathrm{o}_j), \ldots\}$ is determined by searching for the most likely o^*_j, v^* given an o_i over a set of verbs $\mathcal{V}$ extracted from the corpus and the other regions in the image. This is shown in Equation 1.

$$o^*_j, v^* | o_i = \arg\max_{o_j, v} p(o_i) \cdot p(v|o_i) \cdot p(o_j|v, o_i) \qquad (1)$$

We can easily estimate $p(o_i)$, $p(v|o_i)$, and $p(o_j|v,o_i)$ directly from the corpus. If we cannot find an o^*_j, v^* for a region, we back-off to the spatial relationship calculation as defined in PROXIMITY. When we have found the best pairs of regions, we select the most probable pair and generate the first sentence of the description using that pair an template T_1. The second sentence is realised with template T_2 over the subjects in R not used in generating the first sentence. An example of the language generated is:

(2) The man is riding the bike. There is also a car, a road, and trees in the image.

In comparison to PROXIMITY, this model will only describe pairs of regions that have observed relations in the external corpus. The corpus also provides a verb that relates the regions, which produces descriptions that are more in line with human-generated text. However, since noun co-occurrence in the corpus controls which regions can be mentioned in the description, this model will be prone to relating regions simply because their labels occur together frequently in the corpus.

3.3 STRUCTURE

The model STRUCTURE exploits the visual dependency representation of an image to generate language for only the relationships that hold between pairs of regions. It has access to the image regions, the region labels, and the visual dependency representation of an image.

Region–region relationships are generated during a depth-first traversal of the VDR using templates T_1, T_3, T_4, and T_5. The VDR of an image is traversed and language fragments are generated and then combined depending on the number of children of a node in the tree. If a node has only one child then we use T_1 to generate text for the head-child relationship. If a node has more than one child, we need to decide how to order the language generated by the model. We generate sentence fragments using T_4 for each child independently and combine them later. In STRUCTURE, the sentence fragments are sorted by the Euclidean distance of the children from the parent. In order to avoid problematic descriptions such as *"The woman is above the horse is above the field is beside the house"*, we include a special case for when a node has more than one child. In these cases, the nearest region is realized in direct relation to the head using either T_3 (two children) or T_1 (more than two children), and the remaining regions form a separate sentence using T_5. This sorting and combing process would result in *"The woman is above the horse. She is above field and beside the house"* for the case mentioned above.

An example of the type of description that can be generated during a traversal is:

(3) The man is above the bike above the road. There is also a car and trees in the image.

In comparison to PROXIMITY, this model can exploit a representation of an image that encodes the relationships between regions in an image (the VDR). However, it is limited to generating spatial relations, because it cannot predict verbs to relate regions.

3.4 PARALLEL

The model PARALLEL is an extension of STRUCTURE that uses the image descriptions available to predict verbs that relate regions in parent-child relationships in a VDR. At training time it has access to the annotated regions and labels, the visual dependency representations, and the gold-standard image descriptions. Recall from Section 2.1 that the descriptions were dependency-parsed using the parser of McDonald et al. (2005) and alignments were calculated between the nodes in the VDRs and the words in the parsed image descriptions.

We estimate two distributions from the image descriptions using the alignments: $p(verb|o_{head}, o_{child}, rel_{head-child})$ and $p(verb|o_{head}, o_{child})$. The second distribution is used as a backoff when we do not observe the arc label between the regions in the training data. The generation process is similar to that used in STRUCTURE, with two exceptions: (1) it can generate verbs during the generation steps, and (2) when a node has multiple dependents, the sentence fragments are sorted by the probability of the verb associated with them. This sorting step governs which child is in a relationship with its parent. When the model generates text, it only generates a verb for the most probable sentence fragment. The remaining fragments revert back to spatial relationships to avoid generating language that places the subject region in multiple relationships with other regions. An example of the language generated is:

(4) The man is riding the bike on the road. There is also a car and trees in the image.

In comparison to CORPUS, this model generates descriptions in which the relations between the regions determined by the image itself and not by an external corpus. In comparison to PROXIMITY and STRUCTURE, this model generates descriptions that express meaningful relations between the regions and not simple spatial relationships.

4 Image Parsing

The STRUCTURE and PARALLEL models rely on visual dependency representations, but it is unrealistic to assume gold-standard representations will always be available because they are expensive to construct. In this section we describe an image parser that can induce VDRs automatically from region-annotated images, providing the input for the STRUCTURE-PARSED and PARALLEL-PARSED models at test time.

The parser is based on the arc-factored dependency parsing model of McDonald et al. (2005). This model generates a dependency representation by maximizing the score s computed over all edges of the representation. In our notation, $\mathbf{x}_{vis}$ is the set of annotated regions and y_{vis} is a visual dependency representation of the image; (i, j) is a directed arc from node i to node j in $\mathbf{x}_{vis}$, $\mathbf{f}(i, j)$ is a feature representation of the arc (i, j), and $\mathbf{w}$ is a vector of feature weights to be learned by the model. The overall score of a visual dependency representation is:

$$s(\mathbf{x}_{vis}, y_{vis}) = \sum_{(i,j)\in y_{vis}} \mathbf{w} \cdot \mathbf{f}(i, j) \tag{2}$$

The features in the model are defined over region labels in the visual dependency representation as well as the relationship labels. As our dependency representations are unordered, none of the features encode the linear order of region labels, unlike the feature set of the original model. Unigram features describe how likely individual region labels are to appear as either heads or arguments and bigram feature captures which region labels are in head-argument relationships. All features are conjoined with the relationship label.

We evaluate our parser on the 1,023 visual dependency representations from the data set. The evaluation is run over 10 random splits into 80% training, 10% development, and 10% test data.[2] Performance is measured with labelled and unlabelled directed dependency accuracy. The parser achieves 58.2% ± 3.1 labelled accuracy and 65.5% ± 3.3 unlabelled accuracy, significantly better than the baseline of 51.6% ± 2.5 for both labelled and unlabelled accuracy (the baseline was calculated by attaching all image regions to the root node; this is the most frequent form of attachment in our data).

5 Language Generation Experiments

We evaluate the image description models in an automatic setting and with human judgements. In

[2] Different visual dependency representations of the same image are never split between the training and test data.

	Automatic Evaluation				Human Judgements		
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	Grammar	Action	Scene
PARALLEL-PARSED	45.4 ± 2.0	**16.1 ± 0.9**	**6.4 ± 0.7**	**2.7 ± 0.5**	4.2 ± 1.3	**3.3 ± 1.7**	3.5 ± 1.3
PROXIMITY	45.1 ± 0.8	10.2 ± 1.0*	2.1 ± 0.6*	0.4 ± 0.2*	3.7 ± 1.5*	2.1 ± 0.3*	3.0 ± 1.4*
CORPUS	46.1 ± 1.1	12.4 ± 1.3*	3.1 ± 0.8*	0.7 ± 0.3*	4.4 ± 1.1	2.2 ± 1.3*	3.4 ± 1.3
STRUCTURE	40.2 ± 3.0*	11.5 ± 1.2*	3.5 ± 0.5*	0.3 ± 0.1*	4.1 ± 1.4	2.1 ± 1.4*	3.0 ± 1.4*
STRUCTURE-PARSED	41.1 ± 2.1*	12.2 ± 0.9*	3.6 ± 0.4*	0.4 ± 0.2*	4.0 ± 1.4	1.6 ± 1.3*	3.2 ± 1.3
PARALLEL	44.6 ± 3.1	16.0 ± 1.5	6.8 ± 1.0	2.9 ± 0.7	4.5 ± 1.0*	3.4 ± 1.6	3.7 ± 1.3
GOLD	-	-	-	-	4.8 ± 0.4*	4.8 ± 0.6*	4.6 ± 0.7*

Table 4: Automatic evaluation results averaged over 10 random test splits of the data, and human judgements on the median scoring BLEU-4 test split for PARALLEL. We find significant differences ($^\star p < 0.05$) in the descriptions generated by PARALLEL-PARSED compared to models that operate over an unstructured bag of image regions representation. **Bold** means PARALLEL-PARSED is significantly better than PROXIMITY, CORPUS, and STRUCTURE.

the automatic setting, we follow previous work and measure how close the model-generated descriptions are to the gold-standard descriptions using the BLEU metric. Human judgements were collected from Amazon Mechanical Turk.

5.1 Methodology

The task is to produce a description of an image. The PROXIMITY and CORPUS models have access to gold-standard region labels and region boundaries at test time. The STRUCTURE and PARALLEL models have additional access to the visual dependency representation of the image. These representations are either the gold-standard, or in the case of STRUCTURE-PARSED and PARALLEL-PARSED, produced by the image parser described in Section 4. Table 2 provides a reminder of the information the different models have access to at training time.

Our data set of 1,023 image–description–VDR tuples was randomly split into 10 folds of 80% training data, 10% development data, and 10% test data. The results we report are means computed over the 10 splits. The image parser used for models STRUCTURE-PARSED and PARALLEL-PARSED is trained on the gold-standard VDRs of the training splits, and then predicts VDRs on the development and test splits. Significant differences were measured using a one-way ANOVA with PARALLEL-PARSED as the reference[3], with differences between pairs of mean checked with a Tukey HSD test.

5.2 Automatic Evaluation

The model-generated descriptions are compared against the human-written gold-standard descriptions using the smoothed BLEU measure (Lin and Och, 2004). BLEU is commonly used in machine translation experiments to measure the effective overlap between a reference sentence and a proposed translation sentence. Table 4 shows the results on the test data and Figure 4 shows sample outputs for two images. PARALLEL, the model with access to both image structure and aligned image descriptions at training time outperforms all other models on higher-order BLEU measures. One reason for this improvement is that PARALLEL can formulate sentence fragments that relate the subject, a verb, and an object without trying to predict the best object, unlike CORPUS. The probability associated with each fragment generated for nodes with multiple children also tends to lead to a more accurate order of mentioning image regions. It can also be seen that PARALLEL-PARSED remains significantly better than the other models when the VDRs of images are predicted by an image parser, rather than being gold-standard.

[3] Recall that PARALLEL uses gold-standard VDRs and PARALLEL-PARSED uses the output of the image parser described in Section 4.

The weakest results are obtained from a model that relies on the proximity of regions to generate descriptions. PROXIMITY achieves competitive BLEU-1 scores but this is mostly due to it correctly generating region names and determiners. CORPUS is better than PROXIMITY at correctly producing higher-order n-grams than because it has a better model of the region–region relationships in an image. However, it has difficulties guessing the correct verb for a description, as it relies on corpus co-occurrences for this (see the second example in Table 4). STRUCTURE uses the VDR of an image to generate the description, which this leads to an improvement over PROXIMITY on some of the BLEU metrics; however, it is not sufficient to outperform CORPUS.

5.3 Human Judgements

We conducted a human judgement study on Mechanical Turk to complement the automatic evaluation. Workers were paid $0.05 to rate the quality of an image–description pair generated by one of the models using three criteria on a scale from 1 to 5:

1. Grammaticality: give high scores if the description is correct English and doesn't contain any grammatical mistakes.
2. Action: give high scores if the description correctly describes what people are doing in the image.
3. Scene: give high scores if the description correctly describes the rest of the image (background, other objects, etc).

A total of 101 images were used for this evaluation and we obtained five judgments for each image-description pair, resulting in a total of 3,535 judgments. To ensure a fair evaluation, we chose the images from the split of the data that gave median BLEU-4 accuracy for PARALLEL, the best performing model in the automatic evaluations.

The right side of Table 4 shows the mean judgements for each model for across the three evaluation criteria. The gold-standard descriptions elicited judgements around five, and were significantly better than the model outputs on all aspects. Furthermore, all models produce highly grammatical output, with mean ratings of between 3.7 and 4.5. This can be explained by the fact that the models all relied on templates to ensure grammatical output.

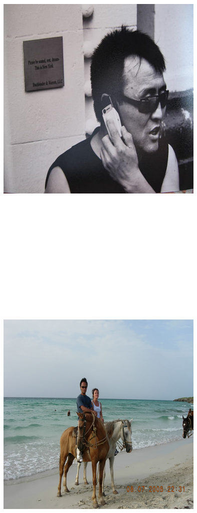
The ratings of the action descriptions reveal the usefulness of structural information. PROXIMITY, CORPUS, and STRUCTURE all perform badly with mean judgements around two, PARALLEL, which uses both image structure and aligned descriptions, significantly outperforms all other models with the exception of PARALLEL-PARSED, which has very similar performance. The fact that PARALLEL and PARALLEL-PARSED perform similarly on all three human measures confirms that automatically parsed VDRs are as useful for image description as gold-standard VDRs.

When we compare the quality of the scene descriptions, we notice that all models perform similarly, around the middle of the scale. This is probably due to the fact that they all have access to gold-standard region labels, which enables them to correctly refer to regions in the scene most of the time. The additional information about the relationships between regions that STRUCTURE and PARALLEL have access to does not improve the quality of the background scene description.

6 Related Work

Previous work on image description can be grouped into three approaches: description-by-retrieval, description using language models, and template-based description. Ordonez et al. (2011), Farhadi et al. (2010), and Kuznetsova et al. (2012) generate descriptions by retrieving the most similar image from a large data set of images paired with descriptions. These approaches are restricted to generating descriptions that are only present in the training set; also, they typically require large amounts of training data and assume images that share similar properties (scene type, objects present) should be described in a similar manner.

Kulkarni et al. (2011) and Li et al. (2011) generate descriptions using n-gram language models trained on a subset of Wikipedia. Both approaches first determine the attributes and relationships between regions in an image as region–preposition–region triples. The disadvantage of relying on region–preposition–region triples is that they cannot distinguish between the main event of the image and the



PROXIMITY	A man is beside a phone. There is also a wall and a sign in the image.
CORPUS	A man is holding a sign. There is also a wall and a phone in the image.
STRUCTURE	A wall is above a wall. A man is beside a sign.
PARALLEL	A man is holding a phone. A wall is beside a sign.
GOLD	A foreign man with sunglasses talking on a cell phone. A large building and a mountain in the background.
PROXIMITY	A beach is above a beach. There are also horses, a woman, and a man in the image.
CORPUS	A woman is outnumbering a man. There are also horses and beaches in the image.
STRUCTURE	A man is beside a woman above a horse. A horse is beside a woman beside a beach.
PARALLEL	A man is riding a horse above a beach. A horse is beside a beach beside a woman.
GOLD	There is a man and women both on horses. They are on a beach during the day.

Figure 4: Some example descriptions produced by PROXIMITY, CORPUS, STRUCTURE and PARALLEL.

background regions. Kulkarni et al. (2011) is closely related to our PROXIMITY baseline.

Yang et al. (2011) fill in a sentence template by selecting the likely objects, verbs, prepositions, and scene types based on a Hidden Markov Model. Verbs are generated by finding the most likely pairing of object labels in an external corpus. This model is closely related to our CORPUS baseline. Mitchell et al. (2012) over-generates syntactically well-formed sentence fragments and then recombines these using a tree-substitution grammar.

Previous research has relied extensively on automatically detecting object regions in an image using state-of-the art object detectors (Felzenszwalb et al., 2010). We use gold-standard region annotations to remove this noisy component from the description generation pipeline, allowing us to focus on the utility of image structure for description generation.

7 Conclusion

In this paper we introduced a novel representation of an image as a set of dependencies over its annotated regions. This *visual dependency representation* encodes which regions are related to each other in an image, and can be used to infer the action or event that is depicted. We found that image description models based on visual dependency representations significantly outperform competing models in both automatic and human evaluations. We showed that visual dependency representations can be induced automatically using a standard dependency parser and that the descriptions generated from the induced representations are as good as the ones generated from gold-standard representations. Future work will focus on improvements to the image parser, on exploring this representation in open-domain data sets, and on using the output of an object detector to obtain a fully automated model.

Acknowledgments

The authors would like to thank M. Lapata and S. Frank for feedback on an earlier draft of the paper and the anonymous reviewers for their feedback. A. M. Enoch, N. Ghahremani-Azghandi, L. S. McAlpine, and K. Tsagkaridis helped annotate the data. The research presented here was supported by the European Research Council under award 203427 Synchronous Linguistic and Visual Processing.

References

Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. 2011. The PASCAL Visual Object Classes Challenge 2011 (VOC2011) Results.

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 *ECCV '10*, 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.

Girish Kulkarni, Visruth Premraj, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C. Berg, and Tamara L. Berg. 2011. Baby talk: Understanding and generating simple image descriptions. In *CVPR '11*, pages 1601–1608, Colorado Springs, Colorado, U.S.A.

Polina Kuznetsova, Vicente Ordonez, Alexander C. Berg, Tamara L. Berg, and Yejin Choi. 2012. Collective Generation of Natural Image Descriptions. In *ACL '12*, pages 359–368, Jeju Island, South Korea.

Siming Li, Girish Kulkarni, Tamara L. Berg, Alexander C. Berg, and Yejin Choi. 2011. Composing simple image descriptions using web-scale n-grams. In *CoNLL '11*, pages 220–228, Portland, Oregon, U.S.A.

Chin-Yew Lin and Franz Josef Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *ACL '04*, pages 605–612, Barcelona, Spain.

Ryan McDonald, Koby Crammer, and Fernando Pereira. 2005. Online large-margin training of dependency parsers. In *ACL '05*, pages 91–98, University of Michigan, U.S.A.

Margaret Mitchell, Jesse Dodge, Amit Goyal, Kota Yamaguchi, Karl Stratos, Alyssa Mensch, Alex Berg, Tamara Berg, and Hal Daum. 2012. Midge : Generating Image Descriptions From Computer Vision Detections. In *EACL '12*, pages 747–756, Avignon, France.

Courtney Napoles, Matthew Gormley, and Benjamin Van Durme. 2012. Annotated Gigaword. In *AKBC-WEKEX Workshop at NAACL-HLT '12*, Montreal, Canada.

Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. 2011. Im2Text: Describing Images Using 1 Million Captioned Photographs. In *NIPS 24*, Granada, Spain.

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

Yezhou Yang, Ching Lik Teo, Hal Daumé III, and Yiannis Aloimonos. 2011. Corpus-Guided Sentence Generation of Natural Images. In *EMNLP '11*, pages 444–454, Edinburgh, Scotland, UK.