Bringing Semantics Into Focus Using Visual Abstraction

C. Lawrence Zitnick
Microsoft Research, Redmond
larryz@microsoft.com

Devi Parikh
Virginia Tech
parikh@vt.edu

Abstract

Relating visual information to its linguistic semantic meaning remains an open and challenging area of research. The semantic meaning of images depends on the presence of objects, their attributes and their relations to other objects. But precisely characterizing this dependence requires extracting complex visual information from an image, which is in general a difficult and yet unsolved problem. In this paper, we propose studying semantic information in abstract images created from collections of clip art. Abstract images provide several advantages. They allow for the direct study of how to infer high-level semantic information, since they remove the reliance on noisy low-level object, attribute and relation detectors, or the tedious hand-labeling of images. Importantly, abstract images also allow the ability to generate sets of semantically similar scenes. Finding analogous sets of semantically similar real images would be nearly impossible. We create 1,002 sets of 10 semantically similar abstract *scenes with corresponding written descriptions. We thoroughly analyze this dataset to discover semantically important features, the relations of words to visual features and methods for measuring semantic similarity.*

1. Introduction

A fundamental goal of computer vision is to discover the semantically meaningful information contained within an image. Images contain a vast amount of knowledge including the presence of various objects, their properties, and their relations to other objects. Even though "an image is worth a thousand words" humans still possess the ability to summarize an image's contents using only one or two sentences. Similarly humans may deem two images as semantically similar, even though the arrangement or even the presence of objects may vary dramatically. Discovering the subset of image specific information that is semantically meaningful remains a challenging area of research.

Numerous works have explored related areas, including predicting the salient locations in an image [17, 26], ranking the relative importance of visible objects [1, 5, 16, 31] and semantically interpreting images [7, 18, 24, 38]. Semantic meaning also relies on the understanding of the attributes of the visible objects [2, 6] and their relations [7, 12]. In common to these works is the desire to understand which visual features and to what degree they are required for semantic understanding. Unfortunately progress in this direction is restricted by our limited ability to automatically extract a diverse and accurate set of visual features from real images.

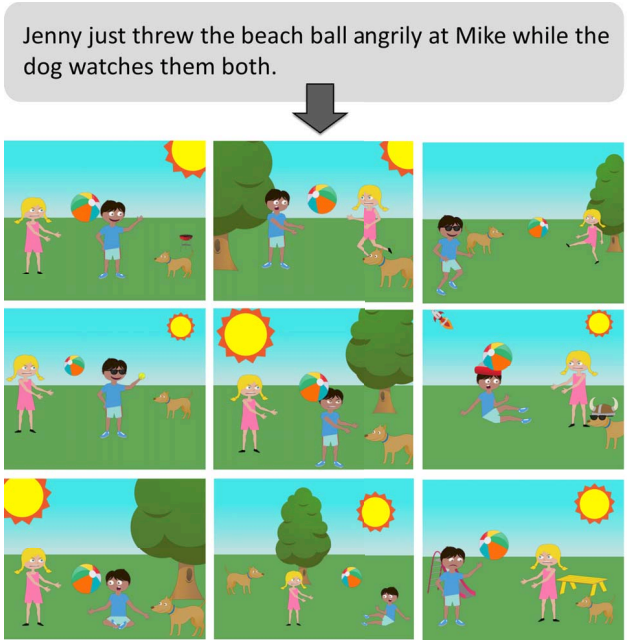

Figure 1. An example set of semantically similar scenes created by human subjects for the same given sentence.

In this paper we pose the question: "Is photorealism necessary for the study of semantic understanding?" In their seminal work, Heider and Simmel [14] demonstrated the ability of humans to endow even simple objects such as triangles and circles with the emotional traits of humans[21]. Similarly, cartoons or comics are highly effective at conveying semantic information without portraying a photorealistic scene. Inspired by these obervations we propose a novel methodology for studying semantic understanding. Unlike traditional approaches that use real images, we hypothesize that the same information can be learned from abstract images rendered from a collection of clip art, as shown in Figure 1. Even with a limited set of clip art, the variety and complexity of semantic information that can be conveyed

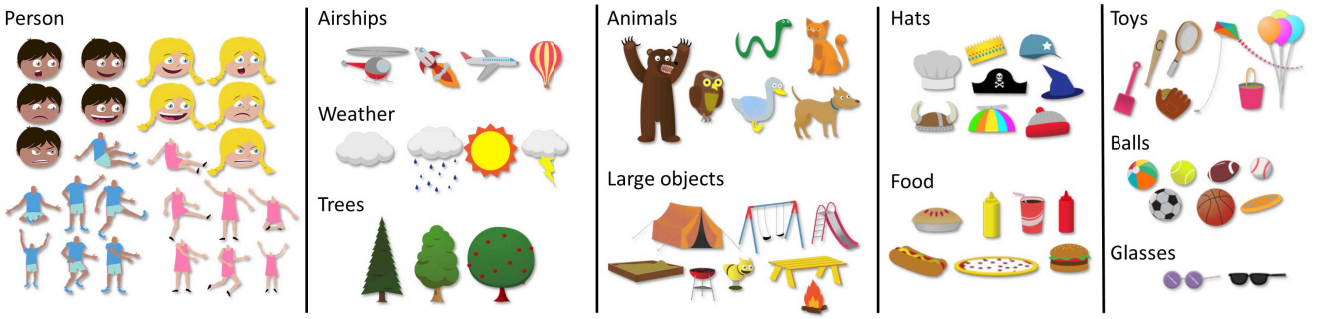

Figure 2. An illustration of the clip art used to create the children (left) and the other available objects (right.)

with their combination is impressive. For instance, clip art can correspond to different attributes of an object, such as a person's pose, facial expression or clothing. Their combination enables an exponential number of potential appearances, Figure 2.

The use of synthetic images provides two main advantages over real images. First, the difficulties in automatically detecting or hand-labeling relevant information in real images can be avoided. Labeling the potentially huge set of objects, their properties and relations in an image is beyond the capabilities of state-of-the-art automatic approaches, and makes hand labeling expensive and tedious. Hand-labeling in many instances is also often ambiguous. Using abstract images, even complex relation information can be easily computed given the relative placement of the clip art, such as "Is the person holding an object?" or "Is the person's or animal's gaze directed towards a specific object?" Second, it is possible to generate different, yet semantically similar scenes. We accomplish this by first asking human subjects to generate novel scenes and corresponding written descriptions. Next, multiple human subjects are asked to generate scenes depicting the same written description without any knowledge of the original scene's appearance. The result is a set of different scenes with similar semantic meaning, as shown in Figure 1. Collecting analogous sets of semantically similar real images would be prohibitively difficult.

Contributions:

- Our main contribution is a new methodology for studying semantic information using abstract images. We envision this to be useful for studying a wide variety of tasks, such as generating semantic descriptions of images, or text-based image search. The dataset and code are publicly available on the author's webpage.
- We measure the mutual information between visual features and the semantic classes to discover which visual features are most semantically meaningful. Our semantic classes are defined using sets of semantically similar scenes depicting the same written description. We show the relative importance of various features, such as the high importance of a person's facial expression or the occurrence of a dog, and the relatively low importance of some spatial relations.
- We compute the relationship between words and visual features. Interestingly, we find the part of speech for a word is related to the type of visual features with which it shares mutual information (*e.g.* prepositions are related to relative position features).
- We analyze the information provided by various types of visual features in predicting semantic similarity. We compute semantically similar nearest neighbors using a metric learning approach [35].

Through our various experiments, we study what aspects of the scenes are semantically important. We hypothesize that by analyzing the set of semantically important features in abstract images, we may better understand what information needs to be gathered for semantic understanding in all types of visual data, including real images.

2. Related work

Numerous papers have explored the semantic understanding of images. Most relevant are those that try to predict a written description of a scene from image features [7, 18, 24, 38]. These methods use a variety of approaches. For instance, methods generating novel sentences rely on the automatic detection of objects [9] and attributes [2, 6, 25], and use language statistics [38] or spatial relationships [18] for verb prediction. Sentences have also been assigned to images by selecting a complete written description from a large set [7, 24]. Works in learning semantic attributes [2, 6, 25] are becoming popular for enabling humans and machines to communicate using natural language. The use of semantic concepts such as scenes and objects has also been shown to be effective for video retrieval [20]. Several datasets of images with multiple sentence descriptions per image exist [11, 28]. However, our dataset has the unique property of having sets of semantically similar images, *i.e.* having multiple images per sentence description. Our scenes are (trivially) fully annotated, unlike previous datasets that have limited visual annotation [11, 28, 36].

Several works have explored visual recognition of different parts of speech. Nouns are the most commonly collected [29, 31] and studied part of speech. Many methods use tagged objects in images to predict important objects

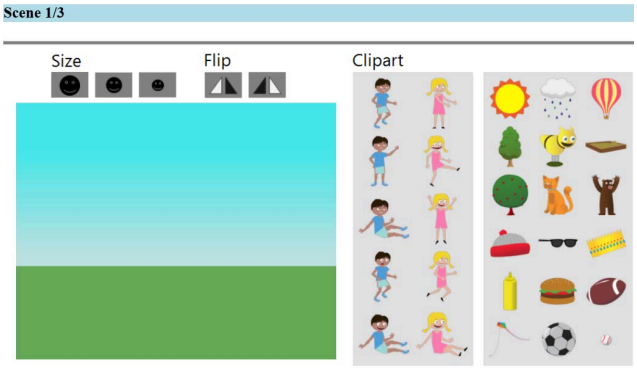

Figure 3. A screenshot of the AMT interface used to create the abstract scenes.

directly from visual features [1, 5, 16, 31], and to study the properties of popular tags [1, 31]. The works on attributes described above includes the use of adjectives as well as nouns relating to parts of objects. Prepositions as well as adjectives are explored in [12] using 19 comparative relationships. Previously, the work of Biederman *et al*. [3] split the set of spatial relationships that can exist in a scene into five unique types. [30] and [39] study the relationships of objects, which typically convey information relating to more active verbs, such as "riding" or "playing". In our work, we explicitly identify which types of visual features are informative for different parts of speech.

3. Generating abstract images

In this section we describe our approach to generating abstract images. The following sections describe various experiments and analysis performed on the dataset.

There are two main concerns when generating a collection of abstract images. First, they should be comprehensive. The images must have a wide variety of objects, actions, relations, etc. Second, they should generalize. The properties learned from the dataset should be applicable to other domains. With this in mind, we choose to create abstract scenes of children playing outside. The actions spanned by children playing cover a wide range, and may involve interactions with a large set of objects. The emotions, actions and interactions between children have certain universal properties. Children also tend to act out "grown-up" scenes, further helping the generalization of the results.

Our goal is to create a set of scenes that are semantically similar. We do this in three stages. First, we ask subjects on Amazon's Mechanical Turk (AMT) to create scenes from a collection of clip art. Next, a new set of subjects are asked to describe the scenes using a one or two sentence description. Finally, semantically similar scenes are generated by asking multiple subjects to create scenes depicting the same written description. We now describe each of these steps in detail.

Initial scene creation: Our scenes are created from a collection of 80 pieces of clip art created by an artist, as shown in Figure 2. Clip art depicting a boy and girl are created from seven different poses and five different facial expressions, resulting in 35 possible combinations for each, Figure 2(left). 56 pieces of clip art represent the other objects in the scene, including trees, toys, hats, animals, etc. The subjects were given five pieces of clip art for both the boy and girl assembled randomly from the different facial expressions and poses. They are also given 18 additional objects. A fixed number of objects were randomly chosen from different categories (toys, food, animals, etc.) to ensure a consistent selection of options. A simple background is used depicting grass and blue sky. The AMT interface is shown in Figure 3. The subjects were instructed to "create an illustration for a children's story book by creating a realistic scene from the clip art below". At least six pieces of clip art were required to be used, and each clip art could only be used once. At most one boy and one girl could be added to the scene. Each piece of clip art could be scaled using three fixed sizes and flipped horizontally. The depth ordering was automatically computed using the type of clip art, e.g. a hat should appear on top of the girl, and using the clip art scale. Subjects created the scenes using a simple drag and drop interface. In all of our experiments, subjects were restricted to United States residents to increase the quality of responses. Example scenes are shown in Figure 1.

Generating scene descriptions: A new set of subjects were asked to describe the scenes. A simple interface was created that showed a single scene, and the subjects were asked to describe the scene using one or two sentences. For those subjects who wished to use proper names in their descriptions, we provided the names "Mike" and "Jenny" for the boy and girl. Descriptions ranged from detailed to more generic. Figure 1 shows an example description.

Generating semantically similar scenes: Finally, we generated sets of semantically similar scenes. For this task, we asked subjects to generate scenes depicting the written descriptions. By having multiple subjects generate scenes for each description, we can create sets of semantically similar scenes. The amount of variability in each set will vary depending on the ambiguity of the sentence description. The same scene generation interface was used as described above with two differences. First, the subjects were given a written description of a scene and asked to create a scene depicting it. Second, the clip art was randomly chosen as above, except we enforced any clip art that was used in the original scene was also included. As a result, on average about 25% of the clip art was from the original scene used to create the written description. It is important to note that it is critical to ensure that objects that are in the written description are available to the subjects generating the new scenes. However this does introduce a bias, since subjects will always have the option of choosing the clip art present in the original scene even if it is not described in the scene description. Thus it is critical that a significant portion of the clip art remains randomly chosen. Clip art that was shown to the original scene creators, but was not used by

them are not enforced to appear.

In total, we generated 1,002 original scenes and descriptions. Ten scenes were generated from each written description, resulting in a total of 10,020 scenes. That is, we have 1,002 sets of 10 scenes that are known to be semantically similar. Figure 1 shows a set of semantically similar scenes. See the author's webpage for additional examples.

4. Semantic importance of visual features

In this section, we examine the relative semantic importance of various scene properties or features. While our results are reported on abstract scenes, we hypothesize that these results are also applicable to other types of visual data, including real images. For instance, the study of abstract scenes may help research in semantic scene understanding in real images by suggesting to researchers which properties are important to reliably detect.

To study the semantic importance of features, we need a quantitative measure of semantic importance. In this paper, we use the mutual information shared between a specified feature and a set of classes representing semantically similar scenes. In our dataset, we have 1002 sets of semantically similar scenes, resulting in 1002 classes. Mutual information (MI) measures how much information the knowledge of either the feature or the class provide of the other. For instance, if the MI between a feature and the classes is small, it indicates that the feature provides minimal information for determining whether scenes are semantically similar. Specifically, if X is the set of feature values, and Y is the set of scene classes,

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log(\frac{p(x,y)}{p(x)p(y)}). \quad (1)$$

Most of our features X are binary valued, while others have continuous values between 0 and 1 that we treat as probabilities.

In many instances, we want to measure the gain in information due to the addition of new features. Many features possess redundant information, such as the knowledge that both a smile and person exist in an image. To measure the amount of information that is gained from a feature X over another feature Z we use the Conditional Mutual Information (CMI),

$$I(X;Y|Z) = \sum_{z \in Z} \sum_{y \in Y} \sum_{x \in X} p(x,y,z) \log(\frac{p(x,y|z)}{p(x|z)p(y|z)}). \quad (2)$$

In the case that we want to condition upon two variables, we compute the CMI for each variable individually and take the minimum value [34]. All scores were computed using 10 random 80% splits of the data. The average standard deviation between splits was 0.002. Next, we describe various sets of features and analyze their semantic importance using Equations (1) and (2).

Occurrence: We begin by analyzing the simple features corresponding to the occurrence of the various objects that may exist in the scene. For real images, this would be the same information that object detectors or classifiers attempt to collect [9]. For occurrence information we use two sets of object types, instance and category. In our dataset, there exist 58 object instances, since we group all of the variations of the boy together in one instance, and similarly for girl. We also created 11 categories by grouping objects of similar type together. These categories, such as people, trees, animals, and food are shown in Figure 2. The ranking of instances and categories based on their MI scores can been seen in Figure 4. Many of the results are intuitive. For instance, objects such as the bear, dog, girl or boy are more semantically meaningful than background objects such as trees or hats. In general, categories of objects have higher MI scores than instances. The semantic importance of an object does not directly depend on how frequently it occurs in the scenes. For instance, people (97.6%) and trees (50.3%) occur frequently but are less semantically important, whereas bears (11.1%) and soccer balls (11.5%) occur less frequently but are important. Interestingly, the individual occurrence of boy and girl have higher scores than the category people. This is most likely caused by the fact that people occur in almost all scenes (97.6%), so the category people is not by itself very informative.

Person attributes: Since the occurrence of the boy and girl are semantically meaningful, it is likely their attributes are also semantically relevant. The boy and girl clip art have five different facial expressions and seven different poses. For automatic detection methods in real images the facial expressions are also typically discretized [8], while poses are represented using a continuous space [37]. We compute the CMI of the person attributes conditioned upon the boy or girl being present. The results are shown in Figure 4. The high scores for both pose and facial expression indicate that human expression and action are important attributes, with expression being slightly higher.

Co-occurrence: Co-occurrence has been shown to be a useful feature for contextual reasoning about scenes [27, 32, 36]. We create features corresponding to the co-occurrence of pairs of objects that occur at least 100 times in our dataset. For our 58 object instances, we found 376 such pairs. We compute CMI over both of the individual objects, Figure 4. Interestingly, features that include combinations of the boy, girl and animals provide significant additional information. Other features such as girl and balloons actually have high MI but low CMI, since balloons almost always occur with the girl in our dataset.

Absolute spatial location: It is known that the position of an object is related to its perceived saliency [33] and can even convey its identity [23]. We measure the position of an object in the image using a Gaussian Mixture Model (GMM) with three components. In addition, a fourth component with uniform probability is used to model outliers.

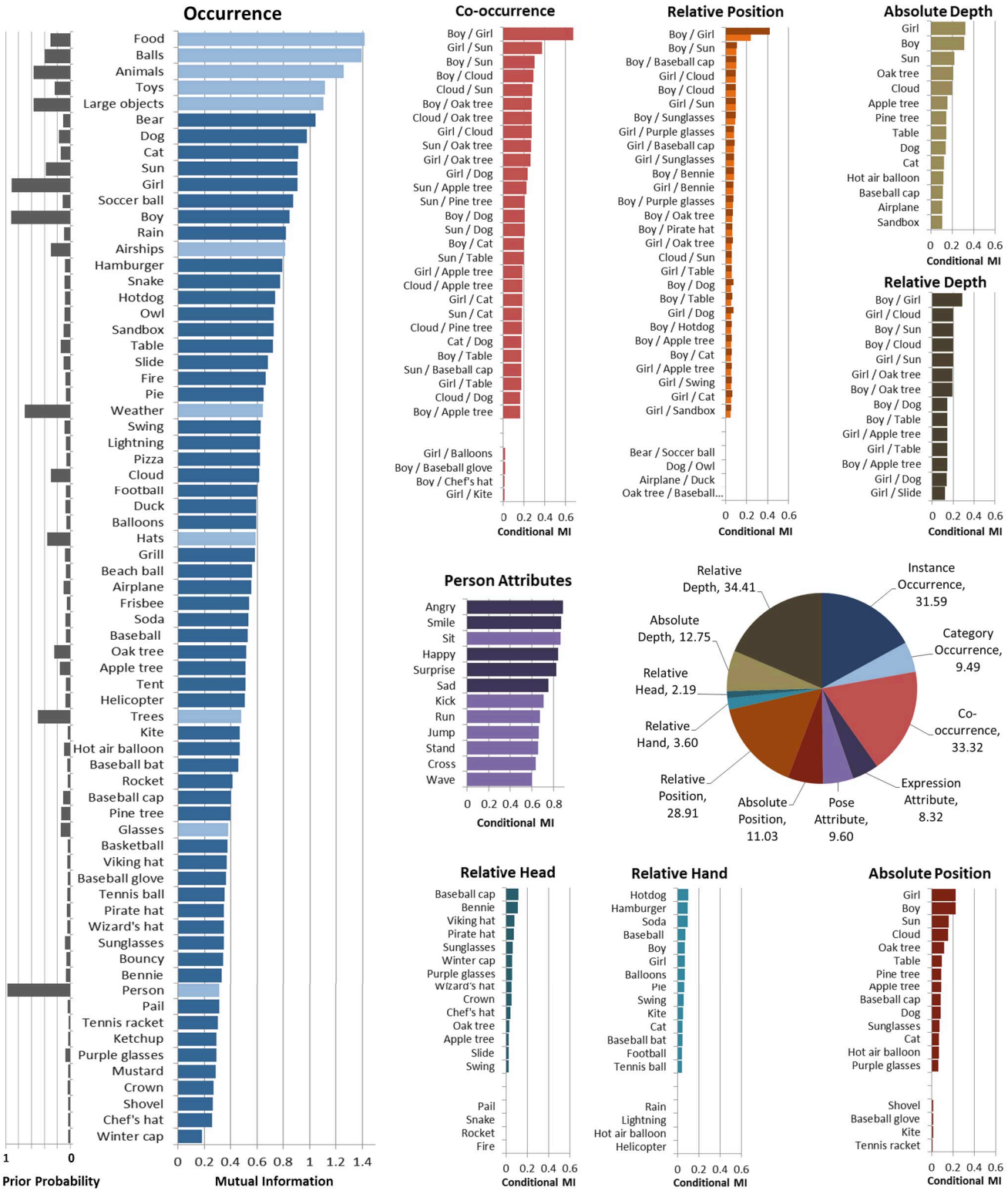

Figure 4. The mutual information measuring the dependence between classes of semantically similar scenes and the (left) occurrence of obejcts, (top) co-occurrence, relative depth and position, (middle) person attributes and (bottom) the position relative to the head and hand, and absolute position. Some mutual information scores are conditioned upon other variables (see text.) The pie chart shows the sum of the mutual information or conditional mutual information scores for all features. The probability of occurrence of each piece of clip art occurring is shown to the left.

Thus each object has four features corresponding to its absolute location in an image. Once again we use the CMI to identify the location features that provide the most additional information given the object's occurrence. Intuitively, the position of the boy and girl provide the most additional information, whereas the location of toys and hats matters less. The additional information provided by the absolute spatial location is also significantly lower than that provided by the features considered so far.

Relative spatial location: The relative spatial location of two objects has been used to provide contextual information for scene understanding [4, 10]. This information also provides additional semantic information over knowledge of just their co-occurrence [3]. For instance, a boy holding a hamburger implies eating, where a hamburger sitting on a table does not. We model relative spatial position using the same 3 component GMM with an outlier component as was used for the absolute spatial model, except the positions are computed relative to one of the objects. The CMI was computed conditioned on the corresponding co-occurrence feature. As shown in Figure 4, the relative positions of the boy and girl provide the most information. Objects worn by the children also provide significant additional information.

One interesting aspect of many objects is that they are oriented either to the left or right. For instance the children may be facing in either direction. To incorporate this information, we computed the same relative spatial positions as before, but we changed the sign of the relative horizontal positions based on whether the reference object was facing left or right. Interestingly, knowledge of whether or not a person's gaze is directed towards an object increases the CMI score. This supports the hypothesis that eye gaze is an important semantic cue.

Finally, we conducted two experiments to measure how much information was gained from knowledge of what a child was holding in their hands or wearing on their head. A single feature using a Gaussian distribution was centered on the children's heads and hands. CMI scores were conditioned on both the object and the boy or girl. The average results for the boy and girl are shown in Figure 4. This does provide some additional information, but not as much as other features. As expected, objects that are typically held in the hand and worn on the head have the highest score.

Depth ordering: The relative 3D location of objects can provide useful information for their detection [13, 15]. The depth ordering of the objects also provides important semantic information. For instance, foreground objects are known to be more salient. Our depth features use both absolute and relative depth information. We create 3 absolute depth features for each depth plane or scale. The relative features compute whether an object is in front, behind or on the same depth plane as another object. The absolute depth features are conditioned on the object appearing while the relative depth features are conditioned on the corresponding pair co-occurring. Surprisingly, as shown in Figure 4, depth

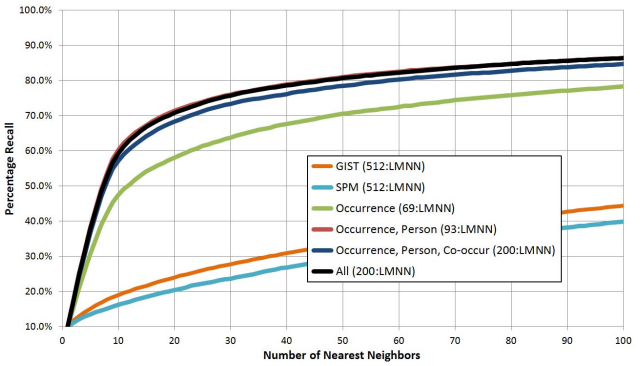

Figure 5. Retrieval results for various feature types. The retrieval accuracy is measured based on the number of correctly retrieved images given a specified number of nearest neighbors.

provides significant information, especially in reference to absolute and relative spatial position.

There are numerous interesting trends present in Figure 4, and we encourage the reader to explore them further. To summarize our results, we computed the sum of the MI or CMI scores for different feature types to estimate the total information provided by them. The pie chart in Figure 4 shows the result. It is interesting that even though there are relatively few occurrence features, they still as a set contain more information than most other features. The person attribute features also contain significant information. Relative spatial and depth features contain similar amounts of information as well, but spread across a much greater number of features. It is worth noting that some of the features contain redundant information, since each was only conditioned upon one or two features. The real amount of information represented by a set of features will be less than the sum of their individual MI or CMI scores.

5. Measuring the semantic similarity of images

The semantic similarity of images is dependent on the various characteristics of an image, such as the object present, their attributes and relations. In this section, we explore the use of visual features for measuring semantic similarity. For ground truth, we assume a set of 10 scenes generated using the same sentence are members of the same semantically similar class, Section 3. We measure semantic similarity using nearest neighbor search, and count the number of nearest neighbors from the same class. We study the recall accuracy using various subsets of our features. In each set, the top 200 features are selected based on MI or CMI score ranking. We compare against low-level image features such as GIST [22] and Spatial Pyramid Models (SPM) [19] since they are familiar baselines in the community. We use a GIST descriptor with 512 dimensions and a 200 visual word SPM reduced to 512 dimensions using PCA. To account for the varying usefulness of features for measuring semantic similarity, we learn a linear warping of the feature space using the LMNN metric learning approach [35] trained on a random 80% of the classes, and tested on

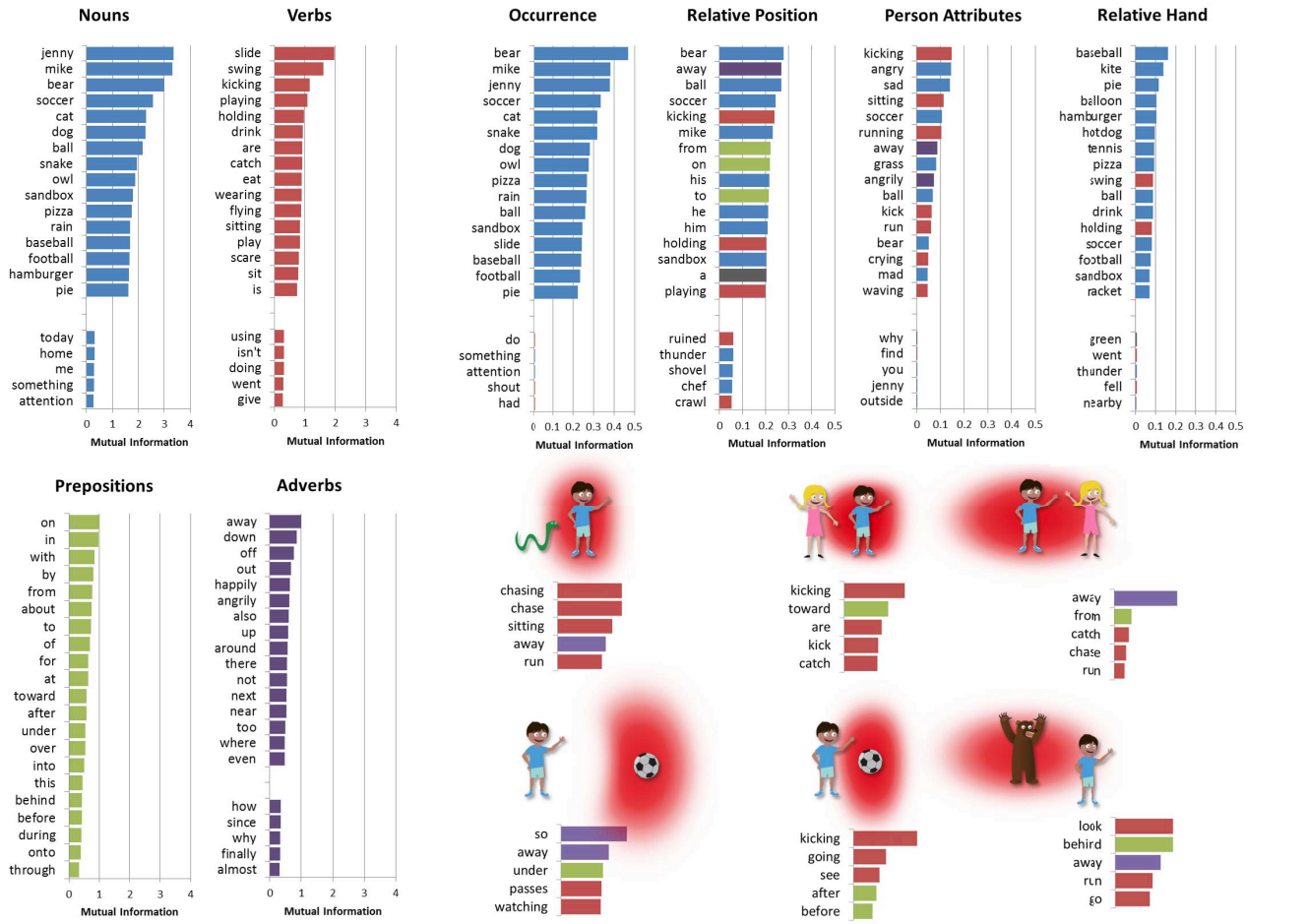

Figure 6. The words with the highest total MI and CMI scores across all features for different part of speech (left). The words with highest total scores across different features types (top-right). Colors indicate the different parts of speech. Top non-nouns for several relative spatial features using object orientation (bottom-right).

the rest. After warping, the nearest neighbors are found using the Euclidean distance.

Figure 5 shows that the low-level features GIST and SPM perform poorly when compared to the semantic (clip art) features. This is not surprising since semantically important information is commonly quite subtle, and scenes with very different object arrangements might be semantically similar. The ability of the semantic features to represent similarity shows close relation to their MI or CMI score in Section 4. For instance the combination of occurrence and person attributes provides a very effective set of features. In fact, occurrence with person attributes has nearly identical results to using the top 200 features overall. This might be partially due to overfitting, since using all features does improve performance on the training dataset.

6. Relating text to visual phenomena

Words convey a variety of meanings. Relating these meanings to actual visual phenomena is a challenging problem. Some words such as nouns, may be easily mapped to the occurrence of objects. However, other words such as verbs, prepositions, adjectives or adverbs may be more difficult. In this section, we study the information shared between words and visual features. In Figure 6, we show for words with different parts of speech the sum of the MI and CMI scores over all visual features. Notice that words with obvious visual meanings (Jenny, kicking) have higher scores, while those with visual ambiguity (something, doing) have lower scores. Since we only study static scenes, words relating to time (before, finally) have low scores.

We also rank words based on different types of visual features in Figure 6. It is interesting that different feature types are informative for different parts of speech. For instance, occurrence features are informative of nouns, while relative position features are predictive of more verbs, adverbs and prepositions. Finally, we show several examples of the most informative non-noun words for different relative spatial position features in Figure 6. Notice how the relative positions and orientations of the clip art can dramatically alter the words with highest score.

7. Discussion

The potential of using abstract images to study the high-level semantic understanding of visual data is especially promising. Abstract images allow for the creation of huge datasets of semantically similar scenes that would be impossible with real images. Furthermore, the dependence on noisy low-level object detections is removed, allowing for the direct study of high-level semantics.

Numerous potential applications exist for semantic datasets using abstract images, which we've only begun to explore in this paper. High-level semantic visual features can be learned or designed that better predict not only nouns, but other more complex phenomena represented by verbs, adverbs and prepositions. If successful, more varied and natural sentences can be generated using visually grounded natural language processing techniques [7, 18, 24, 38].

Finally, we hypothesize that the study of high-level semantic information using abstract scenes will provide insights into methods for semantically understanding real images. Abstract scenes can represent the same complex relationships that exist in natural scenes, and additional datasets may be generated to explore new scenarios or scene types. Future research on high-level semantics will be free to focus on the core problems related to the occurrence and relations between visual phenomena. To simulate detections in real images, artificial noise may be added to the visual features to study the effect of noise on inferring semantic information. Finally by removing the dependence on varying sets of noisy automatic detectors, abstract scenes allow for more direct comparison between competing methods for extraction of semantic information from visual information.

Acknowledgements: We thank Bryan Russell, Lucy Vanderwende, Michel Galley and Luke Zettlemoyer who helped inspire and shape this paper during discussions. This work was supported in part by NSF IIS-1115719.

References

[1] A. Berg, T. Berg, H. Daume, J. Dodge, A. Goyal, X. Han, A. Mensch, M. Mitchell, A. Sood, K. Stratos, et al. Understanding and predicting importance in images. In *CVPR*, 2012.

[2] T. Berg, A. Berg, and J. Shih. Automatic attribute discovery and characterization from noisy web data. *ECCV*, 2010.

[3] I. Biederman, R. Mezzanotte, and J. Rabinowitz. Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive psychology*, 14(2), 1982.

[4] C. Desai, D. Ramanan, and C. Fowlkes. Discriminative models for multi-class object layout. In *ICCV*, 2009.

[5] L. Elazary and L. Itti. Interesting objects are visually salient. *J. of Vision*, 8(3), 2008.

[6] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *CVPR*, 2009.

[7] A. Farhadi, M. Hejrati, M. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth. Every picture tells a story: Generating sentences from images. *ECCV*, 2010.

[8] B. Fasel and J. Luettin. Automatic facial expression analysis: a survey. *Pattern Recognition*, 36(1), 2003.

[9] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *PAMI*, 32(9), 2010.

[10] C. Galleguillos, A. Rabinovich, and S. Belongie. Object categorization using co-occurrence, location and appearance. In *CVPR*, 2008.

[11] M. Grubinger, P. Clough, H. Müller, and T. Deselaers. The iapr tc-12 benchmark: A new evaluation resource for visual information systems. In *Int. Workshop OntoImage*, 2006.

[12] A. Gupta and L. Davis. Beyond nouns: Exploiting prepositions and comparative adjectives for learning visual classifiers. *ECCV*, 2008.

[13] A. Gupta, A. Efros, and M. Hebert. Blocks world revisited: Image understanding using qualitative geometry and mechanics. *ECCV*, 2010.

[14] F. Heider and M. Simmel. An experimental study of apparent behavior. *The American Journal of Psychology*, 1944.

[15] D. Hoiem, A. Efros, and M. Hebert. Putting objects in perspective. In *CVPR*, 2006.

[16] S. Hwang and K. Grauman. Learning the relative importance of objects from tagged images for retrieval and cross-modal search. *IJCV*, 2011.

[17] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *PAMI*, 20(11), 1998.

[18] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. Berg, and T. Berg. Baby talk: Understanding and generating simple image descriptions. In *CVPR*, 2011.

[19] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.

[20] W.-H. Lin and A. Hauptmann. Which thousand words are worth a picture? experiments on video retrieval using a thousand concepts. In *ICME*, 2006.

[21] K. Oatley and N. Yuill. Perception of personal and interpersonal action in a cartoon film. *British J. of Social Psychology*, 24(2), 2011.

[22] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, 42(3), 2001.

[23] A. Oliva, A. Torralba, et al. The role of context in object recognition. *Trends in cognitive sciences*, 11(12), 2007.

[24] V. Ordonez, G. Kulkarni, and T. Berg. Im2text: Describing images using 1 million captioned photographs. In *NIPS*, 2011.

[25] D. Parikh and K. Grauman. Relative attributes. In *ICCV*, 2011.

[26] C. Privitera and L. Stark. Algorithms for defining visual regions-of-interest: Comparison with eye fixations. *PAMI*, 22(9), 2000.

[27] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie. Objects in context. In *ICCV*, 2007.

[28] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier. Collecting image annotations using amazon's mechanical turk. In *NAACL HLT Workshop on Creating Speech and Language Data with Amazon's MT*, 2010.

[29] B. Russell, A. Torralba, K. Murphy, and W. Freeman. Labelme: a database and web-based tool for image annotation. *IJCV*, 2008.

[30] M. Sadeghi and A. Farhadi. Recognition using visual phrases. In *CVPR*, 2011.

[31] M. Spain and P. Perona. Measuring and predicting object importance. *IJCV*, 91(1), 2011.

[32] A. Torralba, K. Murphy, and W. Freeman. Contextual models for object detection using boosted random fields. In *NIPS*, 2004.

[33] P. Tseng, R. Carmi, I. Cameron, D. Munoz, and L. Itti. Quantifying center bias of observers in free viewing of dynamic natural scenes. *Journal of Vision*, 9(7), 2009.

[34] S. Ullman, M. Vidal-Naquet, E. Sali, et al. Visual features of intermediate complexity and their use in classification. *Nature neuroscience*, 5(7), 2002.

[35] K. Weinberger, J. Blitzer, and L. Saul. Distance metric learning for large margin nearest neighbor classification. In *NIPS*, 2006.

[36] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*. IEEE, 2010.

[37] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *CVPR*, 2011.

[38] Y. Yang, C. Teo, H. Daumé III, and Y. Aloimonos. Corpus-guided sentence generation of natural images. In *EMNLP*, 2011.

[39] B. Yao and L. Fei-Fei. Modeling mutual context of object and human pose in human-object interaction activities. In *CVPR*, 2010.