Adopting Abstract Images for Semantic Scene Understanding

C. Lawrence Zitnick, *Member, IEEE,* Ramakrishna Vedantam, and Devi Parikh, *Member, IEEE,*

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* images 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. Finally, we study the relation between the saliency and memorability of objects and their semantic importance.

Index Terms—Semantic Scene Understanding, Linguistic Meaning, Saliency, Memorability, Abstract Images

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 detect salient content and summarize an image 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 salient and semantically meaningful remains a challenging area of research.

Numerous works have explored related areas, including predicting the salient locations in an image [1], [2], ranking the relative importance of visible objects [3], [4], [5], [6] and semantically interpreting images [7], [8], [9], [10]. Semantic meaning also relies on the understanding of the attributes of the visible objects [11], [12] and their relations [13], [7]. 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.

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[15]. Similarly, cartoons or comics are highly effective at conveying semantic information without portraying a photorealistic scene. Inspired by these observations 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

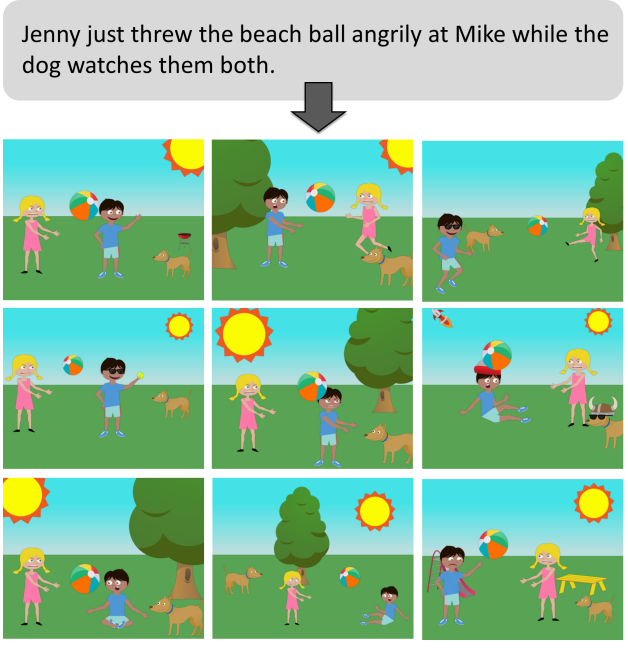

Fig. 1. An example set of semantically similar scenes created by human subjects for the same given sentence.

- *C. Lawrence Zitnick is with Microsoft Research.*
- *Ramakrishna Vedantam and Devi Parikh are with Virginia Tech.*

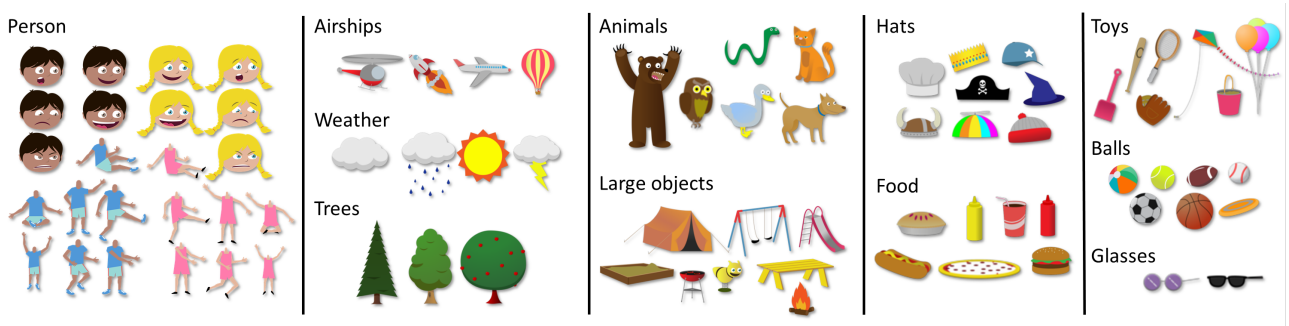

Fig. 2. An illustration of the clip art used to create the children (left) and the other available objects (right.)

Figure 1. Even with a limited set of clip art, the variety and complexity of semantic information that can be conveyed 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.

Related to semantic scene understanding is visual salience and memorability. When describing a scene, many objects, attributes and relations are left unmentioned, while others are so surprising that they are invariably described. Among other factors, saliency and memorability play an active role in whether particular image details are worth mentioning. Previously, many papers have explored low-level cues for finding salient [1], [3], [16] and memorable [17] regions in an image. However, saliency and memorability are also dependent on high-level cues [18], such as the presence and relation of objects, and scene categories. Using abstract images, these high-level cues may be directly studied to gain insights into their contribution to saliency and memorability and how they relate to semantic importance.

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 easily generate related but novel abstract images. One scenario explored in this paper is the generation of 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. Another scenario for using sets of related abstract images is studying object saliency. By analyzing whether human subjects notice the addition or removal of an object from an image we may determine an object's saliency. Numerous other scenarios also exist including object importance, saliency of object relations and scene memorability.

Contributions:

- Our main contribution is a new methodology for studying semantic information and visual salience using abstract images. We envision this to be useful for studying a wide variety of tasks, such as generating semantic descriptions of images, text-based image search, or detecting salient objects. The dataset and code are publicly available on the first 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 [19].
- We study the relation between semantic importance, saliency and memorability of objects. While these concepts are related, they still provide complementary information. Objects of high semantic importance are not always salient or memorable.

Through our various experiments, we study what aspects of the scenes are visually salient and semantically important. We hypothesize that by analyzing semantic importance and high-level visual salience 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

Semantic scene understanding: 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], [8], [9], [10]. These methods use a variety of approaches. For instance, methods generating novel sentences rely on the automatic detection of objects [20] and attributes [11], [12], [21], and use language statistics [9] or spatial relationships [10] for verb prediction. Sentences have also been assigned to images by selecting a complete written description from a large set [7], [8]. Works in learning semantic attributes [11], [12], [21] 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 [22] and grounded natural language generation from video [23], [24], [25], [26]. Several datasets of images with multiple sentence descriptions per image exist [27], [28], [29]. 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 [27], [28], [30].

Linking visual features to different parts of speech: Several works have explored visual recognition of different parts of speech. Nouns are the most commonly collected [31], [5] and studied part of speech. Many methods use tagged objects in images to predict important objects directly from visual features [3], [4], [5], [6], and to study the properties of popular tags [5], [6]. 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 [13] using 19 comparative relationships. Previously, the work of Biederman et al. [32] split the set of spatial relationships that can exist in a scene into five unique types. [33] and [34] 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.

Saliency: Computational models of visual saliency have a rich history [1]. Approaches have ranged from early models which used features inspired from attentional mechanisms [1] to object driven saliency [35]. Human gaze or fixation has been used as a proxy for attention in natural behavior. Recent works have sought to predict these fixations for image regions [36] and objects [37]. Both top down and bottom up models of saliency exist. A comprehensive evaluation of the state of the art is beyond the scope of this paper. We refer the reader to an evaluation [38] for the same. In contrast to most previous works, we are interested in a more high-level notion of saliency that quantifies the extent to which an object is noticed and remembered by human subjects.

Importance: A related notion of importance has also been examined in the community. The order in which people are likely to name objects in an image was studied in [5], while [6] predicted which objects, attributes, and scenes are likely to be described. The order in which the attributes are likely to be named was studied in [39]. Interestingly, [40] and [4] found that the scale and location of an object is related to the order in which a user tags the object in an image. This information may be exploited for improved object detection [40] and image retrieval [4]. In this work we do not explore the order in which objects are mentioned in an image. However, this would be an interesting area for future research using abstract images.

Memorability: People have been shown to have a remarkable ability to remember particular images in long-term memory. [41] demonstrated this ability for images of every day scenes, objects or events, while [42] explored shapes of arbitrary forms. Our memory does not simply include the gist of the picture, but also a detailed representation allowing us to identify which precise image we saw [43], [42]. As most of us would expect, image memorability depends on the user context and is likely to be subject to some inter-subject variability [44]. However, Isola et al. [17] found that despite this expected variability, there is also a large degree of agreement between users. This suggests that there is something intrinsic to images that make some more memorable than others. Isola et al. quantified the memorability of individual images in [17] and then identified semantic characteristics of images that make them memorable in [18]. Works have since looked at modifying the memorability of face images [45] and identifying regions in images that make them memorable [46]. Studying the contribution of different semantic features (e.g. presence, locations, attributes, co-occurrences of objects, etc.) to memorability would go a long way in understanding memorability. Unfortunately, curating or modifying real images to reflect minor perturbations in these semantic features is not feasible. Abstract images provide a promising platform for such in depth analysis. In this paper, we explore whether the presence of certain objects contributes to the memorability of an image. A discussion of different models of memory retrieval [47], [48], [49] and formation [50] are beyond the scope of this paper.

High-level image properties: Many other photographic properties have been studied in the literature such as photo quality [51], saliency [1], attractiveness [52], composition [53], [54], color harmony [55], aesthetics [56] and object importance [5]. In this work we study semantic importance, saliency and memorability of objects, and the relationships of these high-level concepts with each other.

Use of synthetic scenes: Synthetic images and video data have been used to advance computer vision in a variety of ways including evaluating the performance of tracking

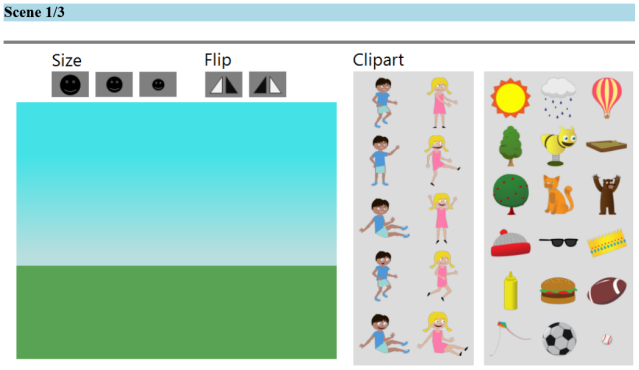

Fig. 3. A screenshot of the AMT interface used to create the abstract scenes.

and surveillance algorithms [57], training classifiers for pedestrian detection [58], human pose estimation [59], learning where to grasp objects [60], evaluating image features for matching patches [61], etc. In this paper, we expand upon Zitnick et al. [62] to further explore the use of abstract images for measuring both object saliency and memorability. The related work of Zitnick et al. [63] models the mapping between language and visual features to automatically synthesize abstract images corresponding to input textual descriptions.

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. We restrict our subjects to be from the United States. 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

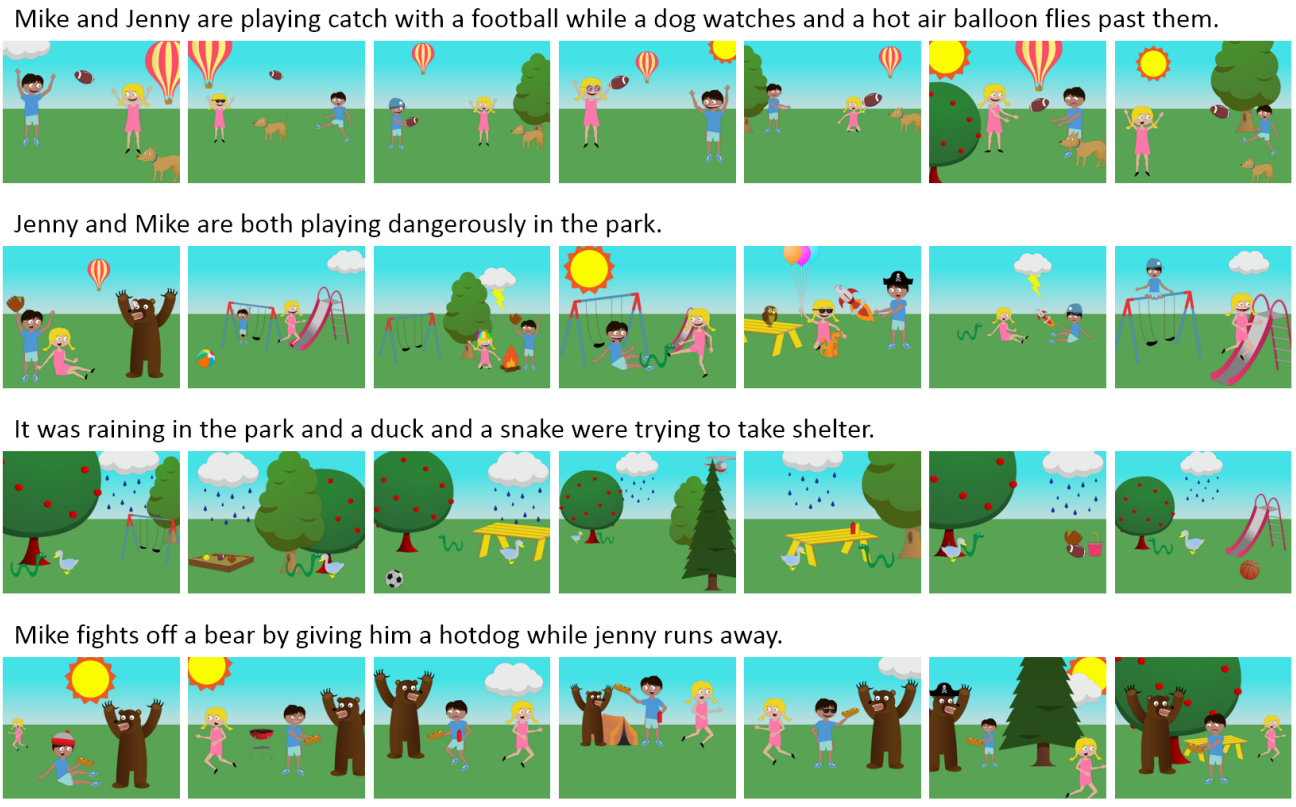

Fig. 4. Example sets of semantically similar scenes. The descriptions may be very specific (top) or more generic (second row.) Notice the variety of scenes that can convey the same semantic description. The presence and locations of objects can change dramatically, while still depicting similar meaning.

Most dependent case (mostly important feature):
values of a feature are same for all scenes in a same class

be semantically similar. Figures 1 and 4 show sets of semantically similar scenes. See the first author's webpage for additional examples.

Independent - p(y|x) = p(y) -> I=0
Completely dependent - p(y|x) = 1 -> I > 0

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,

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

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),

if smile exists, then 'person' must exist regardless of the specific scene.
CMI of 'person' is low

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

In the case that we want to condition upon two variables, we compute the CMI for each variable individually and take the minimum value [64]. 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

all quantitative analyses are based on this MI or CMI index (i.e. a representation of feature importance)

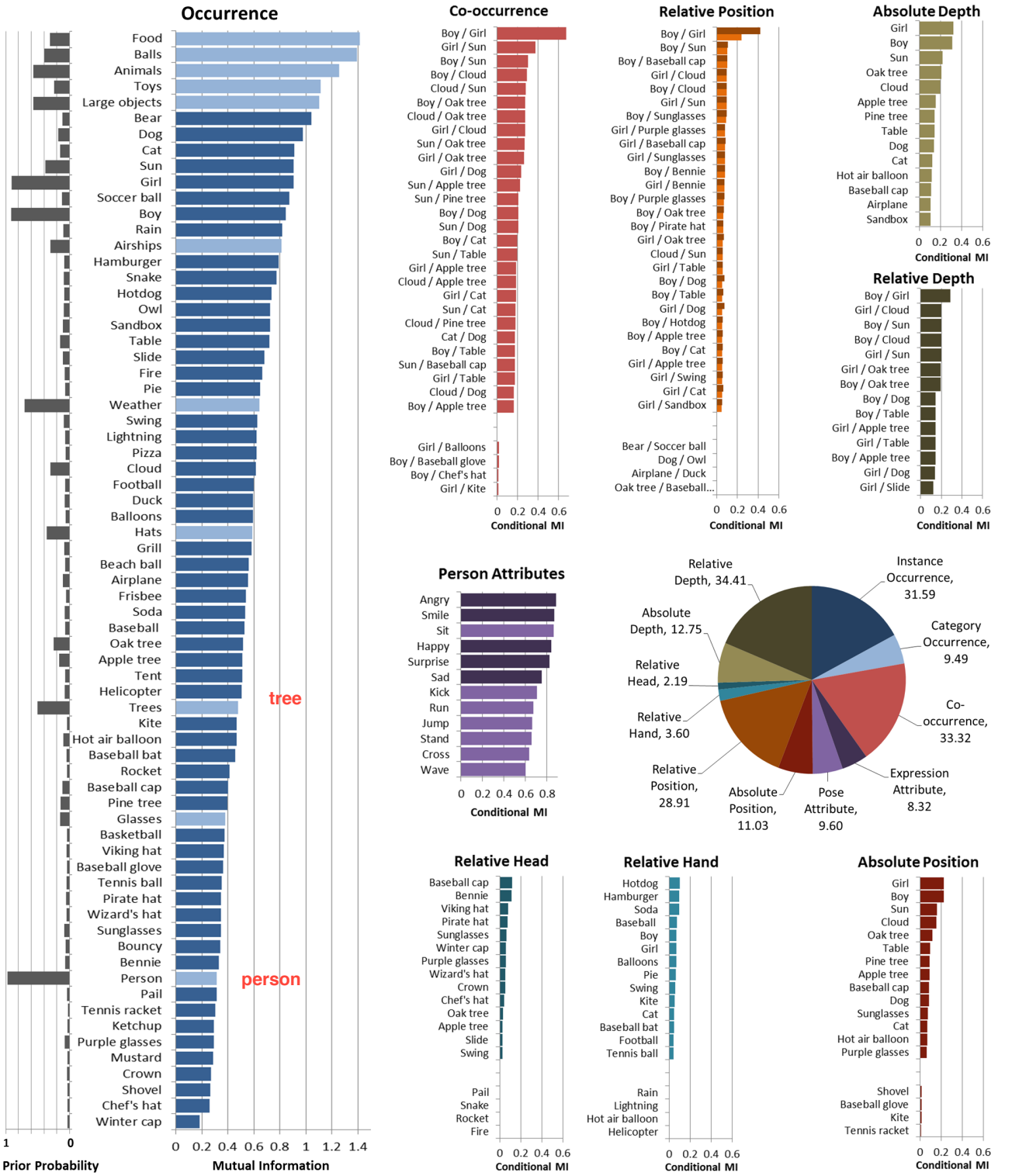

Fig. 5. The mutual information measuring the dependence between classes of semantically similar scenes and the (left) occurrence of obejcts, (top) co-occurrence, relative depth and position, (middle) person attributes and (bottom) the position relative to the head and hand, and absolute position. Some 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.

same information that object detectors or classifiers attempt to collect [20]. 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 5. 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 [65], while poses are represented using a continuous space [66]. We compute the CMI of the person attributes conditioned upon the boy or girl being present. The results are shown in Figure 5. 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 [67], [68], [30]. 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 5. 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 [69] and can even convey its identity [70]. 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. 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.

Positions are continuous variables

Relative spatial location: The relative spatial location of two objects has been used to provide contextual information for scene understanding [71], [72]. This information also provides additional semantic information over knowledge of just their co-occurrence [32]. 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 5, 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 5. 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 [73], [74]. 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 5, depth provides significant information, especially in reference to absolute and relative spatial position.

There are numerous interesting trends present in Figure 5, 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 5 shows the result. It is interesting that even though

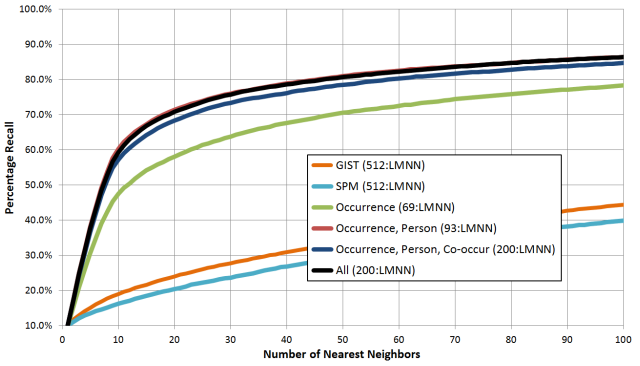

Fig. 6. Retrieval results for various feature types. The retrieval accuracy is measured based on the number of correctly retrieved images given a specified number of nearest neighbors.

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 [75] and Spatial Pyramid Models (SPM) [76] 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 Large Margin Nearest Neighbour (LMNN) metric learning approach [19] trained on a random 80% of the classes, and tested on the rest. After warping, the nearest neighbors are found using the Euclidean distance.

Figure 6 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 7, 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 7. 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 7. Notice how the relative positions and orientations of the clip art can dramatically alter the words with highest score.

7 Object Saliency

manually assessed

In Section 4 we computed the semantic importance of different visual features. In particular, we analyzed how much the occurrence of each object contributes to the semantic meaning of an image. We now study a related but distinct task: How salient is an object in an image? An object is salient if it is noticed and remembered. In other words, if subjects are shown a scene and then the same scene with an object removed, will they notice the difference?

Our experimental set up is similar to that of Isola et al.[17]. We show subjects a series of scenes for one second each. Subjects are instructed to press the 'r' key when they detect a repeated scene. Since our goal is to determine the saliency of an object in a subject's high-level representation of a scene, we introduce a significant delay of approximately four minutes between when the subjects see the target image and its altered version. This is in contrast with tests for visual change detection [77] that show the target and altered image consecutively one after the other. Filler images are shown between the display of the target image and its altered version. Also, filler

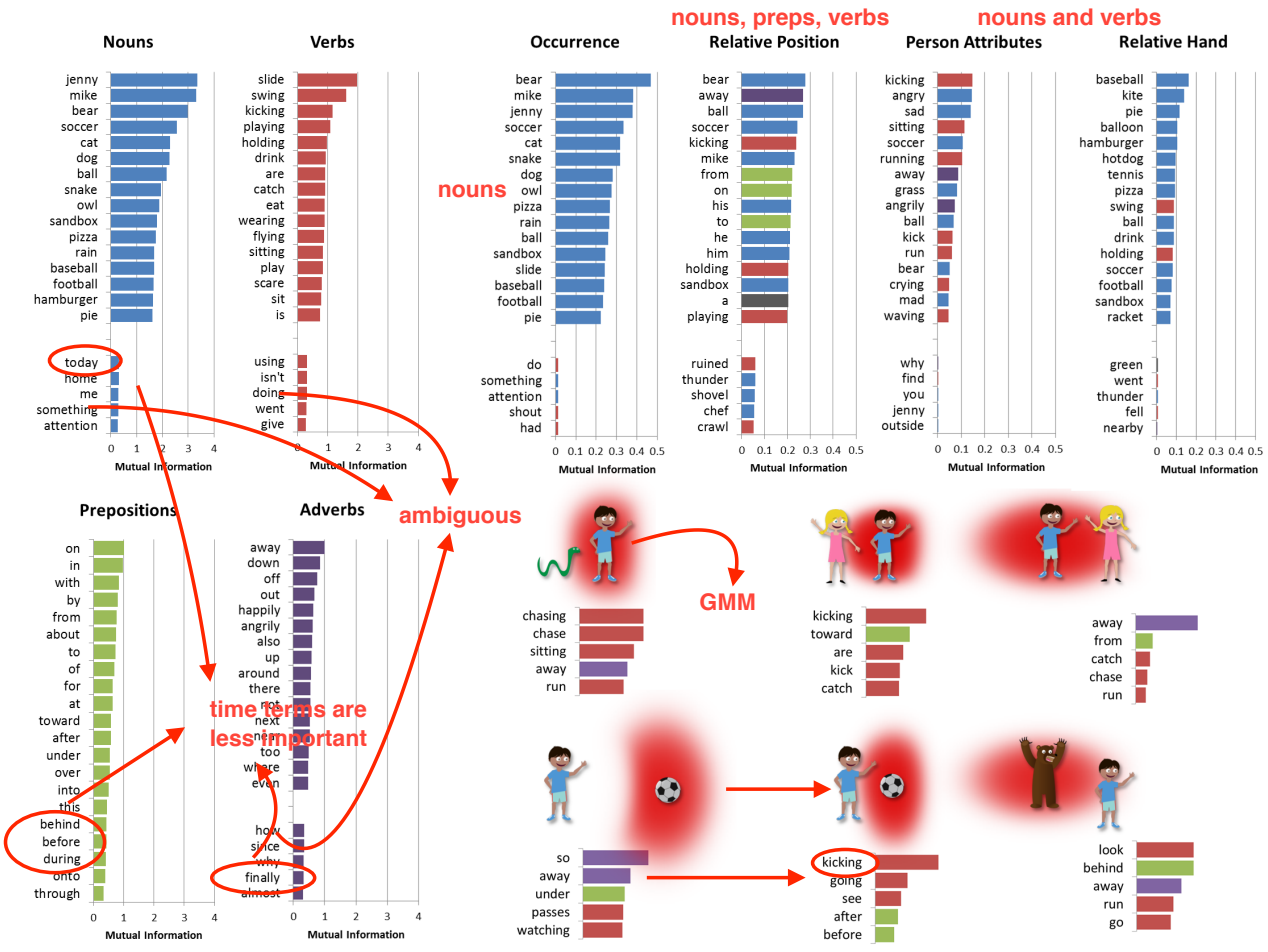

Fig. 7. The words with the highest total MI and CMI scores across all features for different part of speech (left). The words with highest total scores across different features types (top-right). Colors indicate the different parts of speech. Top non-nouns for several relative spatial features using object orientation (bottom-right).

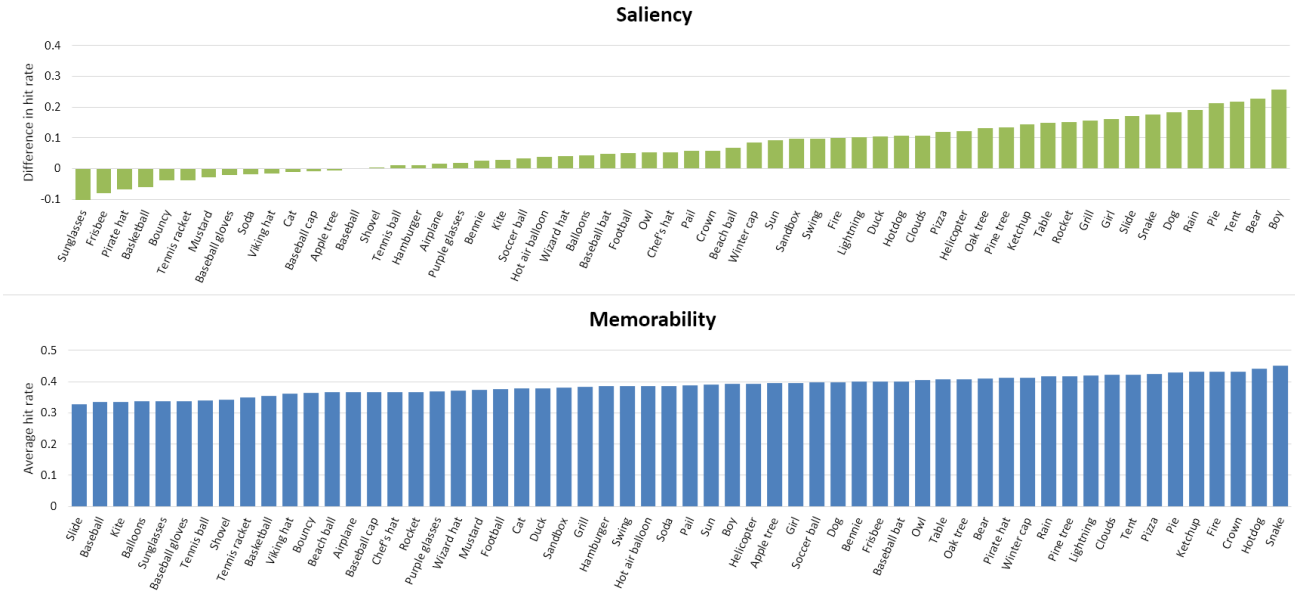

Fig. 8. The saliency (top) and memorability (bottom) for different types of objects. Both are measured using a series of experiments in which subjects are asked to identify scenes they have already seen. The saliency is computed as the difference between the hit rate of intact scenes minus the hit rate of scenes in which an object was removed.

repeats are shown that occur within a shorter duration (2-17 seconds). The filler repeats involve presenting the same filler image again, unaltered. Thus, these serve as vigilance tests to identify workers who are not paying attention or have lost focus.

We randomly selected 4322 scenes as our target scenes. The rest of the scenes in our dataset were used as fillers. One object was chosen to remove from each scene such that each object was removed from an equal number of scenes to the best extent possible. On average each object was removed 148 times (2 subjects on 74 scenes.) A total of 857 unique users participated in our study. For each object we recorded its average hit rate, that is the percentage of target scenes that were labeled as repeating. A scene should have a low hit rate if the object removed was salient. However, a scene may also have a low hit rate if it is simply not memorable. To separate these two confounding factors we ran a second study similar to the first one, but where the target image is repeated without any alterations. The difference in the average hit rates of intact scenes and hit rates of altered scenes gives us an accurate measure of saliency. A high score indicates the object is salient since their hit rate on intact scenes is significantly greater than their hit rate on the altered scenes.

Figure 8(top) shows the difference in hit rates. Interestingly, we find a strong correlation (0.58) between our measure of saliency and semantic importance. We see that objects that have high semantic importance (Figure 5) significantly reduce the hit rate when they are removed, i.e., these objects have high saliency. For instance, if the boy, bear or tent is removed the pair of scenes are not perceived as being the same. It is worth noting that objects that appear smaller or in the background are less salient, such as the sunglasses, frisbee and bouncy. One interesting object is the cat. The cat has high semantic importance, since it is typically described when in the scene. However, since cats typically appear in the background and are not the focus of the scene they have low saliency. Surprisingly for several different objects, their scenes have higher hit rates when the objects are removed, such as sunglasses and frisbee. One possible explanation is the subjects didn't remember these objects upon initially being shown the scene. When viewing an intact scene a second time the objects may be noticed. Since these objects were not remembered, the subjects conclude the scenes are different.

8 Memorability

In the previous section, we determined object saliency by measuring whether an object was noticed and remembered by the subjects. In this section, we explore the memorability of scenes, and whether scenes containing different objects are more likely to be remembered. The memorability of scenes containing different objects is not just dependent on the semantic importance or saliency of the objects present as measured in Sections 4 and 7, but also on the novelty of the objects and their relative locations. Even scenes containing common objects may be highly memorable if the objects are in an unusual configuration. We measure memorability using the same experimental setup as Section 7 using 4322 intact scenes. In Figure 8(bottom) we show the average memorability for scenes containing each object. The objects with high semantic importance are also generally in scenes with higher memorability. The correlation between semantic importance and contribution of an object to memorability was found to be 0.31. However, some objects such as ketchup and crown occur in highly memorable scenes but have lower semantic importance. This may be due to the fact that these objects are often held or worn by the highly salient and semantically important boy and girl and are thus more memorable. Conversely the commonly occurring slide is less memorable while being semantically more important. Due to the numerous factors contributing to memorability, the overall difference between objects was much lower for average scene memorability than for object saliency.

A deeper investigation of the differences between these related but distinct concepts: semantic importance, saliency and memorability is worth pursuing. Abstract scenes provide an appropriate platform to study this. Studying these concepts with regards to other semantic features such as object co-occurrence, relative location and orientation, attributes, etc. is part of future work.

9 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], [8], [9], [10].

One often overlooked aspect of generating written descriptions of images is the need for commonsense knowledge. Determining which aspects of a scene are commonplace or surprising is essential for generating interesting and concise sentences. Many aspects of the scene may also be indirectly implied given commonsense knowledge. For instance, a reader may assume a person is outside if they have knowledge that they are using an umbrella. A promising approach for gathering this commonsense knowledge is through the collection of a large dataset of abstract scenes depicting real world scenes.

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.

ACKNOWLEDGMENTS

We thank Bryan Russell, Lucy Vanderwende, Michel Galley and Luke Zettlemoyer who helped inspire and shape this paper during discussions. We thank John Gruen for his hard work in creating the clip art dataset. We also thank Phillip Isola and Naman Agrawal for the memorability interface. This work was supported in part by NSF IIS-1341772.

REFERENCES

[1] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *PAMI*, 1998.
[2] C. Privitera and L. Stark, "Algorithms for defining visual regions-of-interest: Comparison with eye fixations," *PAMI*, vol. 22, no. 9, 2000.
[3] L. Elazary and L. Itti, "Interesting objects are visually salient," *J. of Vision*, vol. 8, no. 3, 2008.
[4] S. Hwang and K. Grauman, "Learning the relative importance of objects from tagged images for retrieval and cross-modal search," *IJCV*, 2011.
[5] M. Spain and P. Perona, "Measuring and predicting object importance," *IJCV*, vol. 91, no. 1, 2011.
[6] 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.
[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] V. Ordonez, G. Kulkarni, and T. Berg, "Im2text: Describing images using 1 million captioned photographs," in *NIPS*, 2011.
[9] Y. Yang, C. Teo, H. Daumé III, and Y. Aloimonos, "Corpus-guided sentence generation of natural images," in *EMNLP*, 2011.
[10] 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.
[11] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth, "Describing objects by their attributes," in *CVPR*, 2009.
[12] T. Berg, A. Berg, and J. Shih, "Automatic attribute discovery and characterization from noisy web data," *ECCV*, 2010.
[13] A. Gupta and L. Davis, "Beyond nouns: Exploiting prepositions and comparative adjectives for learning visual classifiers," *ECCV*, 2008.
[14] F. Heider and M. Simmel, "An experimental study of apparent behavior," *The American Journal of Psychology*, 1944.
[15] K. Oatley and N. Yuill, "Perception of personal and interpersonal action in a cartoon film," *British J. of Social Psychology*, vol. 24, no. 2, 2011.
[16] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *ICCV*, 2009.
[17] P. Isola, J. Xiao, A. Torralba, and A. Oliva, "What makes an image memorable?" in *CVPR*, 2011.
[18] P. Isola, D. Parikh, A. Torralba, and A. Oliva, "Understanding the intrinsic memorability of images," in *Advances in Neural Information Processing Systems (NIPS)*, 2011.
[19] K. Weinberger, J. Blitzer, and L. Saul, "Distance metric learning for large margin nearest neighbor classification," in *NIPS*, 2006.
[20] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *PAMI*, vol. 32, no. 9, 2010.
[21] D. Parikh and K. Grauman, "Relative attributes," in *ICCV*, 2011.
[22] W.-H. Lin and A. Hauptmann, "Which thousand words are worth a picture? experiments on video retrieval using a thousand concepts," in *ICME*, 2006.
[23] M. Rohrbach, W. Qiu, I. Titov, S. Thater, M. Pinkal, and B. Schiele, "Translating video content to natural language descriptions," in *IEEE International Conference on Computer Vision (ICCV)*, December 2013. [Online]. Available: https://www.d2.mpi-inf.mpg.de/nlg-for-videos-and-images
[24] N. Krishnamoorthy, G. Malkarnenkar, R. J. Mooney, K. Saenko, and S. Guadarrama, "Generating natural-language video descriptions using text-mined knowledge," pp. 10–19, July 2013.
[25] P. Das, C. Xu, R. F. Doell, and J. J. Corso, "A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2013.
[26] H. Yu and J. M. Siskind, "Grounded language learning from video described with sentences," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, vol. 1. Sofia, Bulgaria: Association for Computational Linguistics, 2013, pp. 53–63, best Paper Award. [Online]. Available: http://haonanyu.com/wp-content/uploads/2013/05/yu13.pdf http://haonanyu.com/research/acl2013
[27] 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.
[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] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," *CoRR*, vol. abs/1405.0312, 2014.
[30] 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.
[31] B. Russell, A. Torralba, K. Murphy, and W. Freeman, "Labelme: a database and web-based tool for image annotation," *IJCV*, 2008.
[32] I. Biederman, R. Mezzanotte, and J. Rabinowitz, "Scene perception: Detecting and judging objects undergoing relational violations," *Cognitive psychology*, vol. 14, no. 2, 1982.
[33] M. Sadeghi and A. Farhadi, "Recognition using visual phrases," in *CVPR*, 2011.
[34] B. Yao and L. Fei-Fei, "Modeling mutual context of object and human pose in human-object interaction activities," in *CVPR*, 2010.
[35] W. Einhuser, M. Spain, and P. Perona, "Objects predict fixations better than early saliency," *Journal of Vision*, vol. 8, no. 14, 2008. [Online]. Available: http://www.journalofvision.org/content/8/14/18.abstract
[36] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *IEEE International Conference on Computer Vision (ICCV)*, 2009.
[37] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H.-Y. Shum, "Learning to detect a salient object," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 2, pp. 353–367, 2011.
[38] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 185–207, Jan. 2013. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2012.89
[39] N. Turakhia and D. Parikh, "Attribute dominance: What pops out?" in *ICCV*, 2013.
[40] S. J. Hwang and K. Grauman, "Reading between the lines: Object localization using implicit cues from image tags," *PAMI*, 2012.
[41] L. Standing, "Learning 10,000 pictures," in *Quarterly Journal of Experimental Psychology*, 1973.
[42] I. Rock and P. Englestein, "A study of memory for visual form," in *The American Journal of Psychology*, 1959.
[43] T. F. Brady, T. Konkle, G. A. Alvarez, and A. Oliva, "Visual long-term memory has a massive storage capacity for object details," in *Proceedings of the National Academy of Sciences*, 2008.
[44] R. R. Hunt and J. B. Worthen, "Distinctiveness and memory," in *NY:Oxford Univeristy Press*, 2006.

[45] A. Khosla, W. A. Bainbridge, A. Torralba, and A. Oliva, "Modifying the memorability of face photographs," in *International Conference on Computer Vision (ICCV)*, 2013.
[46] A. Khosla, J. Xiao, A. Torralba, and A. Oliva, "Memorability of image regions," in *Advances in Neural Information Processing Systems (NIPS)*, Lake Tahoe, USA, December 2012.
[47] M. W. Howard and M. J. Kahana, "A distributed representation of temporal context," in *Journal ofMathematical Psychology*, 2001.
[48] G. D. A. Brown, I. Neath, and N. Chater, "A temporal ratio model of memory," in *Psychological Review*, 2007.
[49] R. M. Shiffrin and M. Steyvers, "A model for recognition memory: Rem - retrieving effectively from memory," in *Psychnomic Bulletin and Review*, 1997.
[50] J. L. McClelland, B. L. McNaughton, and R. C. O'Reilly, "Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory," in *Psychological Review*, 1995.
[51] Y. Luo and X. Tang, "Photo and video quality evaluation: Focusing on the subject," in *European Conference on Computer Vision*, 2008.
[52] T. Leyvand, D. Cohen-Or, G. Dror, and D. Lischinski, "Data-driven enhancement of facial attractiveness," *ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH 2008)*, 2008.
[53] B. Gooch, E. Reinhard, C. Moulding, and P. Shirley, "Artistic composition for image creation," in *Eurographics Workshop on Rendering*, 2001.
[54] L. Renjie, C. L. Wolf, and D. Cohen-Or, "Optimizing photo composition," in *Technical report, Tel-Aviv University*, 2010.
[55] D. Cohen-Or, O. Sorkine, R. Gal, T. Leyvand, and Y.-Q. Xu, "Color harmonization," *ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH)*, 2006.
[56] S. Dhar, V. Ordonez, and T. L. Berg, "High level describable attributes for predicting aesthetics and interestingness," in *IEEE Computer Vision and Pattern Recognition*, 2011.
[57] G. R. Taylor, A. J. Chosak, and P. C. Brewer, "Ovvv: Using virtual worlds to design and evaluate surveillance systems," in *CVPR*, 2007.
[58] J. Marin, D. Vazquez, D. Geronimo, and A. Lopez, "Learning appearance in virtual scenarios for pedestrian detection," in *CVPR*, 2007.
[59] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from single depth images," in *CVPR*, 2011.
[60] A. Saxena, J. Driemeyer, and A. Y. Ng, "Robotic grasping of novel objects using vision," *IJRR*, 2008.
[61] B. Keneva, A. Torralba, and W. Freeman, "Evaluation of image features using a photorealistic virtual world," in *ICCV*, 2011.
[62] C. L. Zitnick and D. Parikh, "Bringing semantics into focus using visual abstraction," in *CVPR*, 2013.
[63] C. L. Zitnick, D. Parikh, and L. Vanderwende, "Learning the visual interpretation of sentences," in *ICCV*, 2013.
[64] S. Ullman, M. Vidal-Naquet, E. Sali *et al.*, "Visual features of intermediate complexity and their use in classification," *Nature neuroscience*, vol. 5, no. 7, 2002.
[65] B. Fasel and J. Luettin, "Automatic facial expression analysis: a survey," *Pattern Recognition*, vol. 36, no. 1, 2003.
[66] Y. Yang and D. Ramanan, "Articulated pose estimation with flexible mixtures-of-parts," in *CVPR*, 2011.
[67] A. Torralba, K. Murphy, and W. Freeman, "Contextual models for object detection using boosted random fields," in *NIPS*, 2004.
[68] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie, "Objects in context," in *ICCV*, 2007.
[69] 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*, vol. 9, no. 7, 2009.
[70] A. Oliva and A. Torralba, "The role of context in object recognition," *Trends in cognitive sciences*, vol. 11, no. 12, 2007.
[71] C. Galleguillos, A. Rabinovich, and S. Belongie, "Object categorization using co-occurrence, location and appearance," in *CVPR*, 2008.
[72] C. Desai, D. Ramanan, and C. Fowlkes, "Discriminative models for multi-class object layout," in *ICCV*, 2009.
[73] D. Hoiem, A. Efros, and M. Hebert, "Putting objects in perspective," in *CVPR*, 2006.
[74] A. Gupta, A. Efros, and M. Hebert, "Blocks world revisited: Image understanding using qualitative geometry and mechanics," *ECCV*, 2010.
[75] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *IJCV*, vol. 42, no. 3, 2001.
[76] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *CVPR*, 2006.
[77] R. A. Rensink, "Change detection," *Annual review of psychology*, vol. 53, no. 1, pp. 245–277, 2002.

C. Lawrence Zitnick is a senior researcher in the Interactive Visual Media group at Microsoft Research, and is an affiliate associate professor at the University of Washington. He is interested in a broad range of topics related to visual object recognition. His current interests include object detection, semantically interpreting visual scenes, and the use of human debugging to identify promising areas for future research in object recognition and detection. He developed the PhotoDNA technology used by Microsoft, Facebook and various law enforcement agencies to combat illegal imagery on the web. Previous research topics include computational photography, stereo vision, and image-based rendering. Before joining MSR, he received the PhD degree in robotics from Carnegie Mellon University in 2003. In 1996, he co-invented one of the first commercial portable depth cameras.

Ramakrishna Vedantam is a masters candidate in ECE at Virginia Tech (VT). He is interested in semantic image understanding and structured prediction. As a graduate student, his research focus has been on Scene Understanding. His prior experience involves working on diverse topics in computer vision including evaluation for high level image understanding, face detection, inference in graphical models and sparse 3D reconstruction. He has previously held internship positions at Siemens Corporate Research and Technologies, Bangalore (CTT-IN) and at the Center For Visual Computing (CVC), Ecole Centrale de Paris (ECP) and INRIA-Saclay. He received his bachelors in Electronics and Communication Engineering from International Institute of Information Technology, Hyderabad (IIIT-H) in 2013.

Devi Parikh is an Assistant Professor in the Bradley Department of Electrical and Computer Engineering at Virginia Tech (VT), where she leads the Computer Vision Lab. She is also a member of the Virginia Center for Autonomous Systems (VaCAS) and the VT Discovery Analytics Center (DAC). Her research interests include computer vision, pattern recognition and AI in general and visual recognition problems in particular. She received her M.S. and Ph.D. degrees from the Electrical and Computer Engineering department at Carnegie Mellon University in 2007 and 2009 respectively. She received her B.S. in Electrical and Computer Engineering from Rowan University in 2005. She was a recipient of the Carnegie Mellon Dean's Fellowship, National Science Foundation Graduate Research Fellowship, Outstanding Reviewer Award at CVPR 2012, Marr Best Paper Prize awarded at the International Conference on Computer Vision (ICCV) in 2011, Google Faculty Research Award in 2012, and the 2014 Army Research Office (ARO) Young Investigator Program (YIP) award.