# Bringing Semantics Into Focus Using Visual Abstraction

Zhang, Liqiang

July 8, 2018

## 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 charactering this dependence requires extracting complex visual information from an image, which is in general a difficult and yet unsolved problem. In this paper, the author 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.*

## 1. Introduction

A fundamental goal of computer vision is to discover the semantically meaningful information contained within an image [4]. Images contain a vast amount of knowledge including the presence of various objects [1], 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 content 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 hace explored related areas, including predicting the salient locations in an image, ranking the relative importance of visible objects and semantically interpreting images. Semantic meaning also relies on the understanding of the attributes of the visible and their relations. 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.

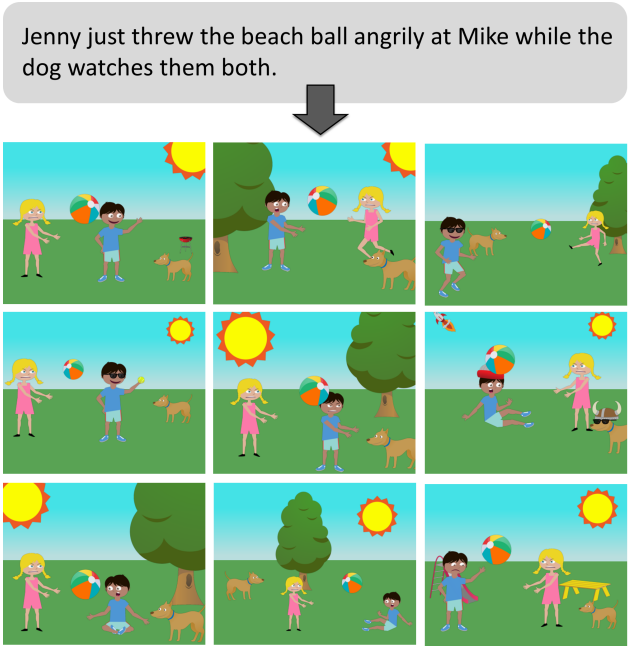


Figure 1. An example set of semantically similar scenes created by human subjects for the same given sentence.

In this paper the author pose the question:"Is photorealism necessary for the study of semantic understanding?" In their seminal work, Heider and Simmel demonstrated the ability of humans to endow even simple objects such as triangles and circles with the emotional traits of humans. Similarly, cartoons or comics are highly effective at conveying semantic information without portraying a photorealistic scene. Unlike traditional approaches that use real images, the same information can be learned from abstract images rendered from a collection of clip art, as shown in Fig. 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 [3]. For instance, clip

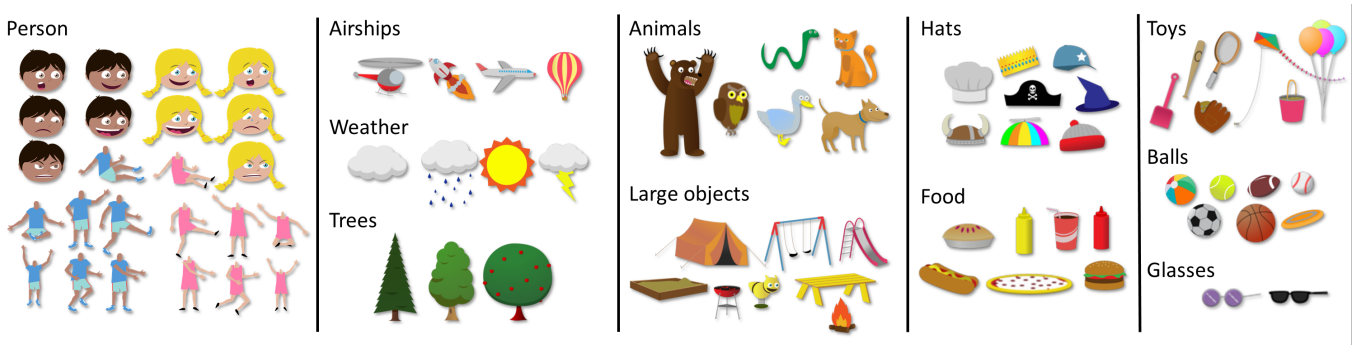


Figure 2. An illustration of the clip art used to create the children (left) and tne other available objects.

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, Fig. 2.

## 2. Semantic importance of visual features

To study the semantic importance of features, it need a quantitative measure of semantic importance. The mutual information can be used to shar between a specified feature and a set of classes representing semantically similar scenes. 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,

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

To measure the amount of information that is gained from a feature $X$ over another feature $Z$ we use the Conditional Mutual In formation (CMI),

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

All scores were computed using 10 random 80% splits of the data. The average standard deviation between splits was 0.002. Next, Eq. 1 and Eq.2 can be used to describe various sets of features and analyze their semantic importance [2].

## References

[1] A. C. Berg, T. L. Berg, I. Daume, Hal, J. Dodge, A. Goyal, X. Han, A. Mensch, M. Mitchell, A. Sood, and K. Stratos. Understanding and predicting importance in images. In *CVPR*, pages 3562–3569, 2012. 1

[2] C. Desai, D. Ramanan, and C. C. Fowlkes. Discriminative models for multi-class object layout. In *ICCV*, pages 229–236, 2010. 2

[3] A. Gupta, A. A. Efros, and M. Hebert. Blocks world revisited: Image understanding using qualitative geometry and mechanics. In *ECCV*, pages 482–496, 2010. 1

[4] C. M. Privitera and L. W. Stark. Algorithms for defining visual regions-of-interest: comparison with eye fixations. *IEEE TPAMI*, 22(9):970–982, 2000. 1