

An Ontology for Generating Descriptions about Natural Outdoor Scenes

Ifeoma Nwogu
University of Rochester,
Rochester, NY 14627

Yingbo Zhou
University at Buffalo, SUNY
Buffalo, NY 14260

Christopher Brown
University of Rochester,
Rochester, NY 14627

Abstract

We present an image ontology useful for generating descriptive texts about highly unconstrained natural outdoor images, taken under many different conditions - lighting, varying viewpoints, etc. The ontology pre-defines the visual content we are interested in describing. Unlike other image description techniques, which tend to be purely object-centric, we utilize a holistic scene ontology for description. The primitive units defined by the ontology are extracted from an image via stochastic processes. Similarly, attributes of the units, also specified by the ontology, are evaluated. Binary and tertiary relationships between relevant primitives are also evaluated. The values, attributes and relationships of the primitive units are combined, based on a pre-defined set of production rules, in such a way as to generate rich, descriptive sentences about the image. Evaluation strategies are implemented to quantitatively test the meaningfulness of the generated sentences. Results indicate that the proposed scene ontology aids in generating highly relevant, naturalistic and meaningful sentences describing natural outdoor images.

1. Introduction and Related Work

Sentences are extremely rich sources of information, both for transmitting and receiving information. Humans can communicate a concise description in the form of a sentence relatively easily. Such descriptions might identify the most interesting objects, what they are doing, and/or where this is happening. These descriptions are rich, accurate, and in good agreement between other humans. They are concise: much is omitted, because humans tend not to mention objects or events that they judge to be less significant. We present an ontology that attempts to identify as much information as possible from an image, although the sentences we generate from the ontology similarly omits redundant information. Although there has been much research in computer vision connecting individual words with structures in pictures, specifically in the area of automatic annotation, there has been less activity in the area of gen-



Figure 1. Below is the sentence generated for the image using our proposed ontology: The image scenery is set in a street in the city; There are at least 6 people in the picture. Someone in the picture has on something wheat colored. The people are standing on the road. The picture background consists mainly of light gray buildings.

erating descriptive texts about images, using a hierarchical knowledge ontology. Some recent work similar to our proposed approach include: (i) Yao *et al.* [14] where the authors use an And-or Graph (AoG) for semi-automatic (human is partly involved) hierarchical image parsing. They also proposed a stochastic image grammar that specifies syntactic or compositional relations as well as and semantic relations between these visual elements. Unfortunately, without an extensive evaluation their process, the complex techniques presented did not clearly indicate how the transition occurred from a parsed image graph to descriptive image sentences. Without evaluation results, it is therefore challenging to compare this technique with others, even qualitatively. (ii) Kulkarni *et al.* [9] present an automatic technique for sentence generation from images that is based on statistics derived from large quantities of text data. Their algorithm defines a conditional random field (CRF) over entities. Although the technique presented was technically sound, the resulting sentences not naturalistic, rather, they

were mainly pairs of entities-with-attributes connected by a preposition. The authors could also have used additional underlying features (such as geometric layouts or depth maps) to obtain more relevant prepositions. Lastly, (iii) **Farhadi et al. [3]** presented a more complex description generation technique where every image is defined by and parsed into a single triplet *object-action-scene* corresponding to a *space of meaning*. Similarly, descriptive sentences, obtained during training, are tagged with the triplet values. A score is obtained by comparing an estimate of meaning obtained from an image to one obtained from a sentence. Although this model is flexible in its design and yielded visually pleasing results, its major drawback was the single triplet-based meaning-space being very limited in its representational power. Also, their space of meaning did not always correspond to human understanding.

Our proposed method, by using a rich image ontology, addresses several of the challenges described above. It is relatively straightforward, where we pre-define the image structure along with the types of sentences we desire to generate. Once all the relevant blocks of the structure are filled in for an image, the production rules can be used to generate the appropriate sentences. Similar to [9], we construct descriptive sentences bottom-up using the image content. Also, our proposed ontology is significantly more representational, as shown by the generated sentences. In describing the images in [9], production rules could have been used to refine the sentences to make them more naturalistic; e.g. *The gray sky is over the gray road*. We use an extensive sets of features to obtain relevant prepositions between entities, and production rules to eliminate any redundant information. Lastly, we show several examples of both good and bad sentence generations, and provide evaluation techniques with results. In comparing our method to the others' work, it is not clear how any of the above described methods would adequately handle a complex image such as in Figure 1.

2. Ontology of Natural Images

2.1. Why develop an ontology [10]?

Gruber [7] defines an ontology as the development of explicit formal specifications of the terms in a domain and relations among them. Formal large-scale ontologies have been developed for the web, ranging from large taxonomies categorizing web sites, to hierarchical categorizations of products for sale along with their features and the characteristics of people who interact with these products (such as on Amazon.com). Standardized ontologies are developed so that domain experts can annotate, organize, and share information in their fields. For example, in medicine, large standardized structured vocabularies of medical terminologies (such as SNOMED) are developed. Ontologies typi-

cally include machine-interpretable definitions of basic concepts in the domain and relations among them. By defining a general scene understanding ontology, we (the AI community) can share a common understanding of the structure of images among people or machine agents. Table 1 provides an overview of the types in the image ontology we propose. Figure 2 shows the semantic structure of our ontology so far.

2.2. Components of the ontology

Entities

Entities act as the basic or primitive units in our ontology. We select these as our primitive units because they are physical, material structures that are readily identified in images, both by humans (through perception) and by machines (through stochastic processes). We define two types of entities, (1) *regions*, which refer to non-countable noun objects. Examples of regions include the sky, foliage, sand, etc.; and (2) *objects* which refer to countable nouns. Some examples of objects are people, cars, boats, etc. Kulkarni et al. [9] make a similar distinction in their work and refer to these entities as *stuff* and *objects*.

Following the paradigm first introduced by Hoiem et al. [8], the entities of a scene can be spatially separated into three main geometric regions - *the sky plane*, *the vertical plane* and *the support or ground plane*. The entities that we define to make up the sky plane are *clear sky* and *clouds*. Entities that make up the verticals include *trees*, *buildings*, *rocks* and *objects*, while those that make up the ground plane include *water*, *grass*, *sand* and *road*. By making such a distinction, we get many *prepositions* in our generated sentences for free. For example, the sky is always spatially above other entities and the verticals are either on/in the support plane. Currently, only entities have attributes in the ontology. The *scene-to-described* is therefore a special type of entity. Though not a material structure, it has its own sets of attributes. Figure 2 shows an overview of how entities are related in the ontology.

Concepts:

We define a concept as a mental picture of a general notion shared by everyone interacting with the ontology. Concepts also have a corresponding language representation. For example, the term *city* invokes a mental picture of a large, urban metropolis. Specifically, the mental associations may be with very different types of city-settings (e.g. the financial district of London versus Chinatown of Los Angeles), but the general notion of a city is shared. Similarly, the term *red* invokes a mental picture of a color at the high wavelength end of the visible spectrum. Specifically, the mental associations may involve colors having very different saturation and intensity values, but the general notion of red is

still shared. It is this notion of concepts that allows to more readily bridge symbolic-semantic (image-to-language) gap. Some concepts in our vocabulary include sunset, garden, city, wheat-colored, red, etc.

Attribute-value pairs:

We specify the properties of entities in an image. Farhadi *et al.* [2] presented an attribute-based framework for describing and naming objects. They learn several attributes of objects, such as rectangular patterns, hairy parts, shiny objects *etc.* They also used the materials of the objects as attributes, such as metallic, wooden *etc.* In addition, we separately introduce a 26-color palette for naming colors on objects.

Relationships:

Although spatial relationships are useful in image descriptions, in order to keep the generated language naturalistic, it is important to discard redundant spatial-relation information in the generated sentences. We attempt to keep such redundancies to a minimum. Relationships that can be described in an image involve the interaction of objects with their surrounding regions, or objects with each other. An example of a tertiary relationship is: *There are six people playing ball on the beach.* Identifying such relationships is still an open area of research in computer vision.

Constraints:

We impose few constraints on the bottom-up process of instantiating and populating the ontology for a particular image. Rather, we implement the stochastic processes to identify as many entities as possible, as well as to assign values to the relevant entity-attributes. Constraints are more strongly implemented during sentence generation instead of during the ontology instantiation.

3. Stochastic Processes for Instantiating the Ontology

The task of instantiating the different concepts identified in the ontology requires us to stochastically extract as much visual content as possible from the images. In this section we describe the different techniques we used to extract visual content from images.

3.1. Image segmentation

We propose a graph based segmentation method where the nodes of the graph are superpixels and adjacent nodes share an edge. The process is initiated with the generation of superpixels where image gradient magnitudes were used as inputs to the watershed algorithm. The resulting superpixels were highly regular in shape and size. The resulting

Type	Meaning
Entity	a physical thing having a distinct, separate, material existence. Entities can therefore be readily identified in an image. For example: <i>{sky, rocks, grass etc.}</i>
Concept	any unit of knowledge associated with both a mental symbol and a corresponding language representation
Formal is-a	<i>is-a</i> is a relationship where one class is a subclass of another, so "D is a B" implies that concept D is a specialization of concept B. For example: <i>{[Road] is a [support].}</i> See Figure 2
Attributes	a specification that defines a property of an entity, usually a name-value pair.
Value restrictions	any range or limitations to the scope of values assigned
Binary relationships	relationships between any two separate entities in the image. For example: <i>{has-a, left-of, on, near, etc.}</i>
Tertiary relationships	relationships between multiple entities, extending pairwise/binary relationships.
Logical constraints	The set of production rules governing the integrity of the ontology. These can include restriction on values allowed by attributes, based on enforcing the binary and tertiary relationships.

Table 1. An overview of the types of information in the image ontology

superpixels were sorted in increasing order of $f(\cdot)$, where f in our case is a single value computed by bit-shifting the average color in a superpixel. The order is traversed once and sorting, the most expensive computation runs in order of $\mathcal{O}(n \log n)$. We extract several features from each superpixel patch, including histogram of hue, saturation, values, and maximum responses from Leung-Malik (LM) filters (texture features). We segment the image by combining the adjacent superpixels, using χ^2 similarity of the feature vectors as the merging predicate, as given in Equation 1. The complete segmentation algorithm is presented in Procedures 1 and 2.

$$\chi^2(f_k, f_l) = \frac{1}{2} \sum_{i=1}^D \frac{[f_k(i) - f_l(i)]^2}{f_k(i) + f_l(i)} \quad (1)$$

f_k, f_l denotes the feature vector of the k-th and l-th node, D is the dimension of the feature vector.

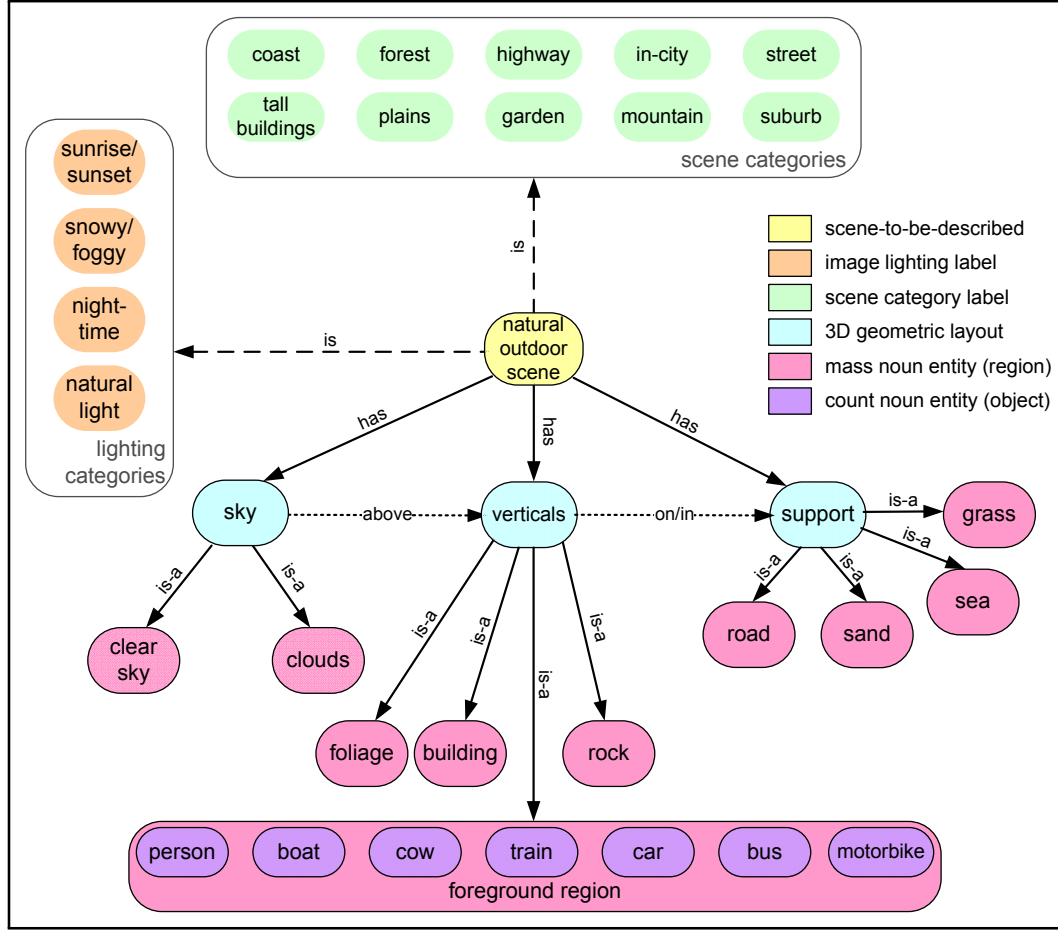


Figure 2. Semantic graph for a natural outdoor image

Procedure 1 SegmentImage(\mathcal{V})

```

Sort all  $v_i \in \mathcal{V}$  in increasing order of mean intensity
for all  $v_i \in \mathcal{V}$  do
  if  $v_i$  has no label then
    Assign  $v_i$  a new label
  end if
  Find the adjacent nodes  $\mathcal{N}$ 
  for all  $n_j \in \mathcal{N}$  do
     $s = \chi^2(\text{Features}(v_i), \text{Features}(n_j))$ 
    if  $s < \text{threshold}$  then
      Assign the label of  $v_i$  to  $n_j$ 
    end if
  end for
end for

```

Scene categorization For scene categorization, we classify the GIST description [12] of the image, performing a one-versus-all test with the rbf kernel on a support vector machine. A 64-dimensional feature vector computed

Procedure 2 Features(v)

```

if  $v$  has been labeled then
  Find the nodes  $\{v'\}$  that have the same label as  $v$ .
  if the number of nodes in  $v' < \text{threshold}$  then
    Calculate color features from all the nodes in  $v'$  and
    store as vector  $F$ .
  else
    Calculate color and texture features from all the
    nodes in  $v'$  and store in  $F$ .
  end if
else
  Calculate color features of node  $v$  and store in  $F$ .
end if
return  $F$ 

```

over the image is used for classification, resulting in one of the classes: $\{\textit{Mountain}, \textit{Woods}, \textit{Beach}, \textit{Street}, \textit{Highway}, \textit{Neighborhood-suburb}, \textit{Neighborhood-city}, \textit{Garden}, \textit{Tall buildings}, \textit{Plains}\}$.

These categories match well to those defined in [12], with the exception of the additional classes. Also, we greatly modified their scene categorization dataset by removing much of the “pure scene” images and included images with mixed scenes (such as a highway by the ocean, or tall buildings surrounded by open plains), for training and testing. In updating the scene training data to more realistic images, the scene classification accuracy rate dropped from 82% as reported originally by the authors to 68% in our dataset.

3.2. Scene illumination using color-location probability maps

Scene illumination is computed by training location probability maps (of opposing color channels) over images belonging to the different illumination classes. A location probability map divides every image in our training set into a fixed number of horizontal slices to coarsely capture its location properties. Each slice is converted to the Red-Green, Blue-Yellow and Intensity channels and 16-bin color histograms are computed for each channel of each slice. The total image feature is therefore a concatenation of the histogram data in a fixed order. SoftMax regression is used to classify each image into *natural light*, *Sunset*, *Night time*, *Snowy*. Examples of R-G and B-Y maps for different illumination scenes are shown in Figure 3. Natural light images often give strong responses to blues in upper locations (presence of skies) and greens in the mid-to-lower locations (trees and grass). Sunset images give strong responses to reds and yellows, especially in the upper regions, Snowy images and night time images are characterized more by their intensity values in all locations in the image, high for snowy images, low for night time images. The details of computing the color conversions are given in [11].

3.3. Spatial geometry recovery and object detections

We use the model by Hoiem *et al.* [8] for estimating surface layouts and the object detectors by Felzenszwalb and Huttenlocher [4] to detect foreground objects and the attributes of the detected objects are then extracted; the objects we detected were *people*, *cars*, *trains*, *motorbikes*, *buses*, *cows* and *boats*.

3.4. Region labeling

Cues for region labels

After segmentation, in order to determine the label of the segments, we extract spatial, color, shape and texture features. Spatial features are composed of the normalized x -, y -means and 10th and 90th percentile of the segment position. The color features for each segment include the mean values from RGB, HSV, $L^*a^*b^*$, XYZ and YCbCr color

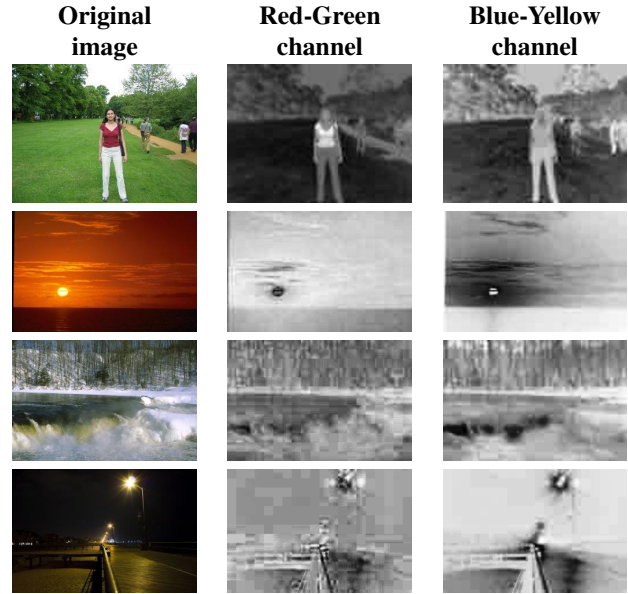


Figure 3. Examples of R-G and B-Y maps for different scene illuminations. The intensity channel is not shown

spaces, as well as the histograms from the hue and saturation channels of the HSV space. The coarse shape features included the number of pixels, and normalized area in an image patch. Mean absolute response, histogram of maximum responses and phase were extracted from LM filter as texture features. From [8] we also borrowed additional holistic features computed over the entire image. These included the vanishing points, estimated horizon, and percentage of nearly parallel pairs of lines. These features were then encoded for each image segment. While validating the learning algorithm for region labeling we observed that (i) Perspective features had strong indications of the building class, since buildings are generally regular in shape, have straight edges and sharp corners, which leads to reliable extraction of long lines at vertical direction. Textures were useful in distinguishing bodies of flowing water from sky. The inclusion of the phase component improved the overall performance.

Labeling regions by fusing cues from the scene class and scene geometry

We implemented a classifier to learn the best features for labeling regions; the classifier consisted of twenty decision trees with each having eight leaf nodes, using the MATLAB `treefit` and subsequently `treeprune` functions. The output of the classifier was further refined by cues obtained from scene classes and scene geometry. We compiled the joint co-occurrence matrices between ground-truth region labels and ground-truth geometry as well as between

the ground-truth region labels and the overall scene classes. Unary potentials were obtained from the classifier outputs for each node segment, the co-occurrence matrices served as binary potentials at the factor nodes between segments and a simple message-passing paradigm updated the geometry and region nodes while scene classes are pegged at their computed values. Additional details on the graphical model and its parameters can be viewed in [11]

4. Generating natural sentences

The goal of our sentence generation is to produce various, simple, well-formed human sentences, describing images of natural outdoor scenes. We used enhanced templates, which contained information needed to generate more complex sentences than just slot filling, to generate the sentences, based on underlying production rules. Enhanced template generation is conceptually straightforward and tailored to the domain, and is therefore of quite good quality. For every visual content theme defined in the ontology, we constructed a pool of sentence templates, so that during generation, one enhanced template is selected per theme. Using the scene classification labels, scene lighting labels, region labels, geometry labels, number and location of objects, and region and object attribute labels, the templates are filled-in. Enhanced templates are modified linguistically to allow for syntax *e.g.* the use of plurals when there is more than one occurring object. The sentence planning/sentence realization tree is shown in Figure 4. Some examples of sentence production rules include:

- Regions labels are included in generated sentences only if the scene illumination is natural lighting. For all other illumination types no regions are included in the sentences (region estimations are inaccurate under all other illuminations)
- Object and regions attributes are included in generated sentences only if the scene illumination is natural lighting. For all other illumination types attributes are not included in the sentences (since attributes rely frequently on colors and textures) The binary relationship between 2 entities is only implemented once to avoid redundancy *e.g.* if we have *the car is on the grass* then we will not include *the grass is beneath the car*. Image attributes are treated as adjectives modifying the entities.

5. Evaluation

Data collection: In order to test our sentence generation paradigm using the proposed ontology, we compiled a new dataset consisting of 709 images. More than 60% of the dataset was obtained from already existing datasets in computer vision: (i) the object annotation dataset by Duygulu

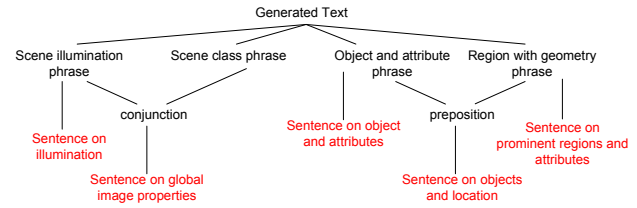


Figure 4. Scheme of how description sentences are planned and realized. Depending on the visual content present in an image, only the relevant sentences are generated to describe that image.

et al. [1]. This dataset was originally obtained from the Corel® 5K dataset; (ii) the scene geometry layout dataset [8]; and (iii) the Stanford region labeling dataset by Gould *et al.* [6]. This dataset also overlapped with the scene geometry dataset. In addition, we also obtained images that were not typically found in computer vision datasets, such as images taken under snowy or foggy conditions, images taken at sunset or at night, with limited lighting. Also included were images taken from atypical viewpoints. Some samples of images and results from our dataset are shown in Table 2. The additional images were obtained from other Corel® datasets, *google images* and *flickr.com*. A total of about 75% of the images in the dataset were taken in daylight while the remaining images are equally distributed over having sunset (or dusk), foggy (or snowy) and night-time illumination.

Once we had completed the collection of the 709 images, we used Amazon’s Mechanical Turk to generate three sets of descriptive sentences for each image. We required the workers to be based in the US and have a working knowledge of the English language. We also checked the sentences reported and rejected sentences with spelling and/or grammatical errors. Details about getting access to the datasets can be obtained by contacting the first or second authors. Our dataset therefore had 3 paragraphs, generated by 3 different users, with a varying number of sentences for each image.

In order to test the meaningfulness of the sentences generated using the proposed ontology, implemented a human-in-the-loop test conducted as follows:

- The evaluation involved 25 testers, each accessing <http://www.zmldesign.com/SUNProject>. The webpage randomly selected one of the 709 images in the dataset and presented its corresponding generated text to the user.
- The auto-selected image along with 19 others (a total of 20 images) were then presented to the tester along with the generated text.
- The tester was encouraged to view all 20 images in detail and select the image that was most appropriately described by the given text.
- A survey page was launched and the tester was encour-

aged to fill the questionnaire. The questions on the survey included: (i) rating how well the description text *explained* the actual image; and (ii) ranking the order of usefulness of the scene context, background regions, objects and attributes in the presented sentences.

* The average meaningful score was 60.34 % *Details on the experimental setup and analysis of results can be found in [11]*

5.1. Discussion of qualitative results obtained

Some sample results are presented in Figure 2. In some cases, as shown in the top three rows, the scene classification, scene illumination, region labeling, and object/attribute detection were in agreement with the actual concepts the images conveyed. In other cases, when the scene illuminations is far from natural lighting, the region labels are meaningless and are not considered during sentence generation. However, if the scene illumination algorithm purports that the scene has natural lighting when it does not, very wrong sentences are generated *e.g.* last row of the results. In other cases such as the 7th row (the wedding image), although the labels are inaccurate, the generated sentences are still moderate. In summary, all components of the ontology have to work correctly together to generate true, meaningful descriptions.

6. Conclusions and Future work

In conclusion, we have presented an ontology defining the visual content of images, useful for generating naturalistic sentences that are meaningful to humans. We performed quantitative tests that indicated that over 60% of the time, the generated sentences at least fairly explained the content in the presented images.

Currently, our visual content knowledge extraction is an amalgamation of many different stochastic processes. Although the advantage to this approach is its modularity (where changes can occur in one module without affecting the entire system), it will also be useful to develop a more unified stochastic framework where the underlying processes can share common features. Going forward, it will be useful to study what aspects of a scene yields the most information and under what circumstances. Our test website provides the opportunity for such an evaluation. Also from our survey results, we intend to study the statistical distributions of the scene aspects that users found most useful and correlate these with the underlying image content.

Also, we empirically observed that the correctness of the object and attribute detections in images greatly affects the meaningfulness of the generated text. Therefore, one main goal going forward is to improve the quality of object detection. In order improve the generated sentences, it will be







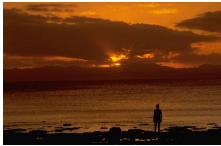



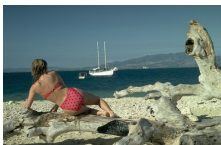






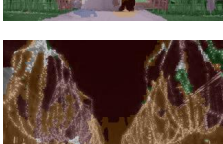
important to include additional information which is obvious humans. For example, the inclusion of gender and age specificity when describing people-in-images, inclusion of commonly known places, such as the London tower or Eiffel Tower in France, *etc.* The goal going forward is so glean as much visual content as possible from an image (we will need to expand the concepts in the current ontology) and statistically generate sentences with only the relevant information.

Acknowledgements

Nwogu is funded by the NSF Computing Innovations Fellowship under the NSF subaward No. CIF-384.

References

- [1] P. Duygulu, K. Barnard, J. F. G. de Freitas, and D. A. Forsyth. Object Recognition as Machine Translation: Learning a Lexicon for a Fixed Image Vocabulary. In *ECCV*, pages 97–112, 2002. 6
- [2] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing Objects by their Attributes. In *CVPR*, 2009. 3
- [3] A. Farhadi, S. M. M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. A. Forsyth. Every Picture Tells a Story: Generating Sentences from Images. In *ECCV*, pages 15–29, 2010. 2
- [4] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object Detection with Discriminatively Trained Part-Based Models. *TPAMI*, 32:1627–1645, 2010. 5
- [5] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient graph-based image segmentation. *IJCV*, 59:167–181, 2004.
- [6] S. Gould, R. Fulton, and D. Koller. Decomposing a scene into geometric and semantically consistent regions. In *ICCV*, 2009. 6
- [7] T. R. Gruber. A Translation Approach to Portable Ontology Specification. *Knowledge Acq.*, pages 199–220, 1993. 2
- [8] D. Hoiem, A. A. Efros, and M. Hebert. Recovering Surface Layout from an Image. *IJCV*, 75(1):151–172, 2007. 2, 5, 6
- [9] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. Berg. Baby Talk: Understanding and Generating Image Descriptions. In *CVPR*, 2011. 1, 2
- [10] N. F. Noy and D. L. McGuinness. Ontology development 101: A guide to creating your first ontology. 2
- [11] Nwogu, I. and Zhou, Y. and Brown C. DISCO: Describing Images using Scene Contexts and Objects. In *AAAI*, 2011. 5, 6, 7
- [12] A. Oliva and A. Torralba. Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope. *IJCV*, 42(3):145–175, 2001. 4, 5
- [13] E. Reiter and R. Dale. *Building Natural Language Generation Systems*. Cambridge University Press, New York, NY, USA, 2000.
- [14] Z. Yao, X. Yang, L. Lin, M. W. Lee, and S. Zhu. I2t: Image parsing to text description. *Proc. of IEEE*, 98(8):1485–1508, 2010. 1

Actual image	Auto-text	Human-text	Labeled regions
	The picture shows a scene... in a street in the city; There is a light gray car in the picture. The car is on the road. The picture background consists mainly of light gray buildings.	A narrow street that has two cars parked on it, with large white buildings on three sides of the street.	
	The picture shows a scene... with a view of tall buildings; The picture has a person in something purplish-pink. The picture background consists mainly of light gray buildings. There's quite a large expanse of dark gray trees in the scenery.	The building is white. There are many people gathered around the white building. The white building is very tall. The white building has many windows.	
	The image scenery is set with a view of the water; The picture has a person in something dark gray. There is also a motorbike in the picture. The person is standing on the road. The motorbike is on the road. The light gray road are quite prominent in the scenery. The picture background consists mainly of dark green trees.	A guy on a red and black motorcycle wearing a red, white, and black jacket. He also has on blue jeans and white sneakers. His helmet is on the ground next to him.	
	This scene is at sunset or dusk... with a view of the water; There is a person in the picture.	A lone figure watching the sunset over a lake. One person watching the sunset over the mountains.	
	The picture shows a night scene... in a street in the city; There is a car in the picture.	The image is in the night time. There are a few cars parked on the side of the street. The street lights are on in this picture.	
	This scene shows a snowy day... a suburban area; The picture has a person in something brown.	This is an image of a woman looking out to the ocean. A ship is in the ocean. The woman is wearing a pink bikini.	
	This scene shows a snowy day... in a street in the city;	A shoveled sidewalk in front of brick buildings. A shoveled sidewalk next to a bus stop and a truck driving in the street.	
	The image scenery is set in a garden; There are at least 2 people in the picture. Someone in the picture has on something black. The people are standing on the grass. The picture background consists mainly of dark gray buildings. The light gray trees are quite prominent in the scenery.	This is an image of a man and woman getting married. The woman is dressed in white. The man is dressed in black. They have white flowers and decorations all around them.	
	The picture shows a scene... in a street in the city; There's quite a large expanse of black buildings in the scenery.	Blue lights in the heart of a city. A street lined with strings of blue lights.	

sky
 tree
 road
 grass
 water
 bldg
 mntn
 fg obj
 sand
 cloud

Table 2. Some results of generating sentences from the ontology; column 1: original images, column 2: labeled regions, column 3:auto-generated sentences, column 4: human description