Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence

DISCO: Describing Images Using Scene Contexts and Objects

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

In this paper, we propose a bottom-up approach to generating short descriptive sentences from images, to enhance scene understanding. We demonstrate automatic methods for mapping the visual content in an image to natural spoken or written language. We also introduce a human-in-the-loop evaluation strategy that quantitatively captures the meaningfulness of the generated sentences. We recorded a correctness rate of 60.34% when human users were asked to judge the meaningfulness of the sentences generated from relatively challenging images. Also, our automatic methods compared well with the state-of-the-art techniques for the related computer vision tasks.

1 Introduction

The ability to have a machine correctly parse, recognize and communicate naturally about the visual world around is a significant advancement in machine intelligence, which tacitly is an augmentation of the human intellect. For real-life scenes, 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.

Similar to how humans use sentences to communicate, we present naturalistic machine generated sentences, which when presented to humans can be used to identify specific scenes being described. As an example, the text below is a sentence generated from our system, *DISCO (Describing Images using Scene Contexts and Objects)*. The reader is encouraged to select one of the four images presented in Figure 1 that is best described by the sentences below.

```
The image scenery is set with a view of tall buildings;
There are at least 5 people in the picture. Someone in
the picture has on something bluish-purple. The people
are standing on the ground. The light gray buildings are
quite prominent in the scenery.
```

Our descriptions apply to general, all-purpose outdoor images. These are significantly more challenging to describe than images in a specific domain such as street scenes, maritime scenes etc, since there are very few domain-specific constraints that can be applied to aid scene understanding. Included in our data set are images taken in natural light, at dusk and at dawn, in foggy weather, in snow storms and at night. Also included in the data set are images taken at different viewpoints. Unlike the images in many of the scene categorization benchmark datasets, which typically only contain "pure scenes", our dataset contains single images that simultaneously belong to multiple scene categories, and taken at multiple viewpoints. In many of these images, everything potentially co-occurs with everything else. Some examples from our test dataset are shown in Figure 1. The complete results of vision-to-language descriptor can be accessed at
`http://www.zmldesign.com/SUnProject` [1].

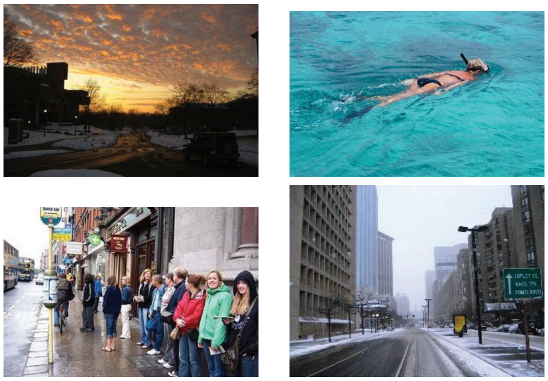


Figure 1: Examples of images from our sentence generation dataset.

1.1 Physically grounded application scenarios

Scenario 1: Travel Companion Bot. An example of an application of the technique we propose is a scene-conversation-system deployed in a car, having the ability

[1] Web address is case-sensitive. The reader is encouraged to test and fill in richer descriptions to help advance this research

to converse with its driver concerning its surrounding outdoor environment. *Knight Rider* was an American television series that originally ran from 1982 to 1986, involving a sapient, artificial intelligence-based, talking car, *KITT*. Implementing the technology to realize some of the mechanisms manifested in the fictitious robot, KITT, such as being able to "understand and communicate about physical visual surrounding" would be an advancement in artificial intelligence. Since cars do not remain only within a specific scene category, but can be driven within any outdoor scene such as for a picturesque family vacation trip, to the beach, within the city, etc., having an all-purpose scene-independent image description generator is not far-fetched in such a context. Such a car/robot would need to understood its surroundings and be able to communicate in a natural language with its driver and passengers. The conversation could be high-level general scene descriptions, or could be semantically "zoomed-in", to be specific about an object or an event in the scene.

Scenario 2 - Safety patrol Bot. This scenario involves autonomous robots patrolling highways, beaches, country roads, oil rigs, etc., with a variety of sensors to capture videos and sound. By providing such robots with the ability to speak in a natural language, a large amount of visual information can be compressed via language and transferred at significantly lower bandwidth, to a human supervisor, and potentially to other patrol bots in the area.

DISCO is currently in the early stages of realizing such physically grounded robots, by providing rich, descriptive sentences about physical environments.

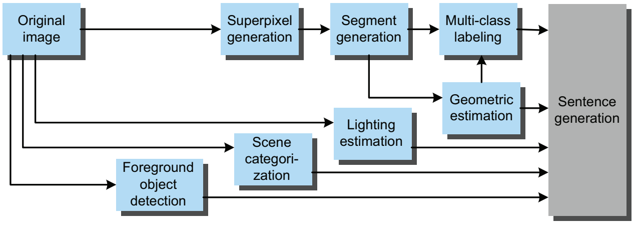


Figure 2: End-to-end process flow of our visual content extraction and sentence generation approach

1.2 Overview of our approach

Figure 2 presents an overview of our visual content extraction and sentence generation process which is initiated with an original input image presented to our system *DISCO*. The global image properties - scene categorization and lighting estimation are first performed on the entire image. Image segments are then computed and presented to the multi-class labeling and geometric layout estimation modules in order to generate the specific details of the image. We pre-defined an ontology of outdoor images consisting of placeholders for scene categories, foreground objects, lighting conditions, background regions and surface layouts. Sentences about the image are constructed from the specific instances of the data types in the ontology, using a forward chaining inference engine.

1.3 Contributions

We summarize our primary contributions as (i) an automated end-to-end technique for generating natural language descriptions from the visual content in images, that can be potentially useful in different real-world applications; (ii) defining a novel computer vision problem of classifying image illumination; (iii) a novel real-time technique for selecting optimal regions in, and parsing ad-hoc outdoor images. We show that our technique compares well with the state-of-the-art methods on public datasets; and (iv) a quantitative evaluation technique with humans-in-the-loop, for measuring the accuracy and *meaningfulness* of generated image descriptions. Unlike typical tests in computer vision which are initiated by humans (e.g. a person presents an input image and the system returns the image segments or edges), this test is initiated by the computer and a human judge is expected to respond to measure of the meaningfulness of the proposed image description; The advantage of this technique is its quantitative measure of sentence meaningfulness.

2 Related Work

Sentence generation from images touches on many facets of computer vision that have previously been treated as separate problems. The first of these, image segmentation, is primarily concerned with perceptually grouping image parts from lower-level primitive image features, without any knowledge of the image content, in order to obtain meaningful higher-level structures. Image segmentation has been addressed extensively in the computer vision literature, including thresholding techniques, edge-based methods, connectivity-preserving relaxation methods (such as active contours and level set techniques) and region-based techniques. In this study we are only concerned with region-based segmentation which involves partitioning an image into connected regions by grouping neighboring pixels of similar feature values (or cues). These include intensity, color, texture, motion, etc. Many of the advancements in region-based image segmentation are therefore either improvements in cue selection (also referred to as model order selection) such as (Ren, Fowlkes, and Malik 2005), and improvements in optimal cue combination (Alpert et al. 2007), (Shi and Malik 2000), (Comaniciu and Meer 2002). A study by (Rabinovich et al. 2006) proposed a framework that combined both techniques and as an output presented the user with a continuum of *stable* segmentations, rather than just one solution. A stable clustering is one that is repeatable over several perturbations.

Multi-class image labeling, the next facet we review, has been successfully addressed by a number of research works; (Gould, Fulton, and Koller 2009) present an extensive overview of region-based segmentation and labeling techniques in the "background and related work" section of their paper. The goal here in image labeling is to assign every image pixel a single class label. Usually the technique involves constructing a Conditional Random Field (CRF) over either pixels or small coherent regions, superpixels. An appearance based unary potential and a pairwise smoothing potential is defined to encourage neighboring pixels to take the same class label. (Gould, Fulton, and Koller 2009) propose a labeling technique using multiple stable segmentations as input and define an energy minimization technique over several aspects of the scene, including geometric and depth estimations. Other works such as by (Li, Socher, and Li 2009), (Tu et al. 2005) have developed unified methods to simultaneously segment, label and/or categorize images.

Lastly, in the area of sentence generation from general image data (not domain specific images such as sports) there is little reported research. (Farhadi et al. 2010) generated sentences from images, using a model that assumed a space of *Meanings*. A score is obtained by comparing an estimate of meaning obtained from an image to one obtained from a sentence. Each estimate of meaning comes from a discriminative procedure that is learned from training data. Although this model is very flexible in its design and yielded visually pleasing results, it is not obvious how well their defined meanings space corresponds to human understanding, the ultimate goal in our proposed technique. As the authors stated in the paper, their "underlying estimate of meaning is impoverished". The technique though, shows tremendous promise for text-based, query-driven image retrieval and high-level object detection from images, the other side of our human-to-computer communication goal, unaddressed in this paper.

Another example of research in sentence generation from images was reported by (Gupta et al. 2009) who generated narrations relating to a sports event in videos, using a compositional model of AND-OR graphs. Their research showed very encouraging results in the specific domain and the structures of expected events were known beforehand, which greatly enhanced their sentence generation as well as the evaluation of results. (Yao et al. 2010) also generated sentences from temporal sequences in pre-specified domains (surveillance in an urban traffic scene and in a maritime scene), also using an AND-OR graph over a visual knowledge representation. Although they demonstrated their algorithms within these specific domains in their paper, no evaluation for sentence correctness was provided.

3 Visual Content Extraction

Image segmentation Our image segmentation paradigm is initiated with the generation of superpixels $j = 1, \dots N$. Image gradient magnitudes were used as inputs to the watershed algorithm (as implemented in Matlab image processing toolkit) and the resulting superpixels are highly regular in shape and size. Our segmentation algorithm first sorts the resulting superpixels 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 $\mathcal{O}(n \log n)$ Each region in the sorting order will be combined with its neighbors based on a merging predicate given by:

$$P(R_i, R_{\mathcal{N}(i)}) = \begin{cases} true & \text{iff} \quad |\bar{R}_i - \bar{R}_{N(i)}| \geq \bar{T}, \\ false & \text{otherwise} \end{cases}$$

where R_i is a region (containing one or more already merged superpixels) and $R_{\mathcal{N}(i)}$ is an unmerged neighbor of R_i; $\bar{R}_i$ is a vector representing the aggregated cues in region R_i. The aggregation in each cue is described in the list below. $\bar{T}$ is a vector of threshold values, where each dimension of $\bar{T}$ is a cue threshold value, learned from training data. The operator $|\cdot|$ represents the χ^2 distance.

The feature list used in our merging predicate are:

- HSV colors: represented by histogram distributions. The aggregate value is the histogram re-distribution of the merged region.
- Texture: represented by the mean absolute responses and histogram responses of 4-scale, 6-orientation LM filters. To compute aggregate values, the unnormalized histograms are summed. After the merge is completed, the texture histograms are normalized.

Multiple segmentations can be encouraged by varying the threshold values so that stable segmentations can be extracted. But in our experiments, by using multiple segmentations, our pixelwise accuracy dropped from about 79% to 70%, hence we opted to use only a single segmentation.

Scene categorization For scene categorization, we classify the GIST description (Oliva and Torralba 2001) of the image, performing a one-versus-all test with the rbf kernel on a support vector machine. A 64-dimensional feature vector is computed over the image is used for classification, resulting in a class C, where $C \in \{$*Mountain, Woods, Beach, Street, Highway, Neighborhood-suburb, Neighborhood-city, Garden, Tall buildings, Plains*$\}$.

These categories match well to those defined in (Oliva and Torralba 2001), 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 accuracy rate dropped from 82% as reported by the authors to 68%.

Scene illumination using location probability maps The scene illumination is computed by training location probability maps 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 color space and 16-bin color histograms are computed for each channel in each slice. The total image feature is therefore a concatenation of the histogram data in a fixed order. Auto-logistic regression is used for classification, resulting in a class L, where $L \in$ *natural light, Sunset, Nighttime, Foggy*. The R-G and B-Y conversion is given by:

$$\begin{pmatrix} O_1 \\ O_2 \\ O_3 \end{pmatrix} = \begin{pmatrix} 5.2 & -4.9 & 1.7 \\ -1.3 & -1.9 & 3.0 \\ 1.0 & 0.0 & 1.0 \end{pmatrix} \begin{pmatrix} 0.7R \\ 1.1G \\ 1.0B \end{pmatrix}$$

O_1 is the R-G channel while O_2 is the B-Y channel. After testing several other color spaces, we found this to be the most discriminant of scene illumination, with its sensitivity to red-green and blue-yellow separately.

Spatial geometry recovery and object detections We use the model by (Hoiem, Efros, and Hebert 2007) for estimating surface layouts and the object detectors by (Felzenszwalb et al. 2010) to detect foreground objects. The objects-of-interest in *DISCO*are *people, cars, trains, motorbikes, buses, cows and boats*

3.1 Fusing scene categories, multi-class region labels and geometric locations

Our region labeling aims to assign a label to each segments in the image. We assume that the segment resolution is sufficient to capture all regions of interest in the image. Each segment is labeled by initially assigning a class to it, based solely on its appearance. Node potentials are based on the pre-computed appearance features, and they include the absolute differences of the mean RGB, L*a*b*, HSV, texture filter responses and x-y location values of the segment centroid, the symmetrized Kullback-Leibler divergence of the color and texture histograms and perspective differences. The ensuing region labels are assigned at the segment level as R_j.

Our fusion model seeks to improve the overall estimated scene class, region and geometry labels and object detection. We accomplish this by incorporating both scene-based and object based context into our final inference.

Figure 3 shows the simplified graphical model of our fusion technique, where the graph consists of two segments S_m and S_n with feature vectors with features X_m and X_n respectively. The joint distribution of the model can be written as:

$$P(C, R_{1\dots N}, G_{1\dots N}|X_{1\dots N}) = \prod_{i=1}^{N} \psi_i \prod_{j=1}^{N} \phi_j \qquad (1)$$

where $\psi_i = \psi(R_i, G_i, X_i)$ is a clique potential, and $\phi_j = \phi(R_j, C)$.

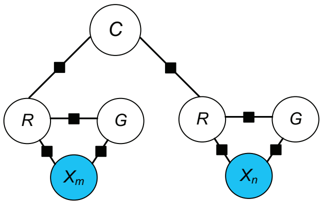


Figure 3: The simplified fusion graphical model showing the factorization for an image with only two segments

The 3-clique is factorized into pairwise potentials for ease of computations.

$$\psi(R_i, G_i, X_i) \propto p(R_i|X_i)\dot{p}(G_i|X_i)\dot{p}(R_i, G_i) \qquad (2)$$

The first two elements on the RHS of Equation 2 were pre-computed during the independent label classification. $p(R_i, G_i)$ is represented by a Region-Location matrix computed during training where the number of co-occurrences of regions and geometrical labels were counted, recorded and normalized. Similarly, $\phi(R_j, C)$ is derived by computing the co-occurrences of regions and scene categories.

4 Sentence Generation

Given the visual content extracted from an image, it is possible to generate sentences about the image using forward chaining production rules. The system consists of the pre-defined set of rules, a working memory to store the temporary data and the forward chaining inference.

```
Input: original image, empty string-finalstr, $C$, $L$, $\{R\}$, $\{G\}$,
        $\{bboxes\}$
Output: Populated string-finalstr
Generate a static color palette of 26 colors;
Randomly select from a list of scene lighting sentences
Populate the scene lighting sentence with token $L$;
Append new sentence to finalstr
Randomly select from a list of scene category sentences
Populate the scene category sentence with token $C$;
Append new sentence to finalstr
foreach object category $j$ do
    foreach bboxes $i$ do
        if Score(i,j) $\geq$ threshold then
            if $L$ = 'Natural lighting' or 'Foggy' then
                Compute foreground object color;
                Populate object sentence with number and
                colors of objects $j$;
            end
            else
                Populate object sentence with number of
                objects $j$;
            end
            Append new sentence to finalstr;
        end
    end
end
if $L$ = 'Natural lighting' or 'Foggy' then
    Populate templates with region information in the ground,
    vertical and sky planes from $\{R\}$ and $\{G\}$;
    Compute background region colors;
    Append new sentences to finalstr;
end
```

Algorithm 1: Sentence generation from image content

The output of the visual content extraction process for an image include: (i) the most likely scene category C selected from the set of nine category labels; (ii) the scene lighting category L selected from the set of four labels; (iii) the set of background regions, $\{R\}$; (iv) the set of geometric regions$\{G\}$; and (v) the set of foreground object bounding boxes *bboxes*, for seven object categories. Color estimations for objects or regions are performed in the L*a*b* space using Euclidean distances between the object/region median value and the pre-defined 26-color palette.

The sentence generation is summarized in Algorithm 1

5 Experiments and Results

Due to the multi-faceted nature of our work, we conducted multiple experiments to evaluate the performance of the intermediary computer vision sub-processes, and one final human-in-the-loop experiment to evaluate the meaningfulness of the generated sentences.

Scene Categorization The scene categorization algorithm was initially presented by Torralba et. al. (Oliva and Torralba 2001), and the "pure scene" dataset presented has since been extended by several other researchers including (Li, Socher, and Li 2009). We discarded the indoor labels and used only the outdoor ones. We extended the training/validation dataset by introducing several real-life images containing multiple scenes and objects. Figure 4 shows the results of running the algorithm both on the "pure scene" dataset, and on the more realistic diverse dataset that we extended. Although the cross-validation accuracy dropped from about 80% to 68% after the dataset was extended, its performance on our final more realistic testing dataset improved.

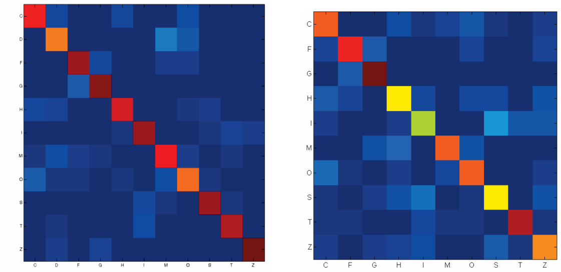


Figure 4: The confusion matrix on the left shows our algorithm using the pure scene data, while the right is the same algorithm on the more realistic dataset.

Image illumination: Image illumination training was performed on a dataset containing 324 training images. Of these, 20% were images taken at night, 20% were images taken at dusk or sunset, 10% were images taken in snowy or foggy conditions and the remaining 50% were images taken under natural lighting conditions. Our illumination classifying algorithm was tested via 5-fold cross validation testing, and resulted in an accuracy of **78%**. Since there is currently no publicly available illumination benchmark dataset to compare against, we will make our training dataset publicly available for testing.

Image segmentation and labeling: Our image segmentation-labeling algorithm was tested on the Stanford dataset containing 715 images of outdoor scenes compiled by (Gould, Fulton, and Koller 2009); we performed the same tests in the paper, a 5-fold cross-validation, with the dataset randomly split into 572 training and 143 test images for each fold. Our pixel-level accuracy was **79.84% (0.77)** - standard deviation are given in parenthesis. A baseline standard pixelwise CRF model gave a mean pixel accuracy of 74.30% (0.80) and the region-based energy model by (Gould, Fulton, and Koller 2009) had an accuracy of 76.40% (1.22). It is worth noting that their energy minimization technique can take as long as ten minutes to segment one image, while our algorithm runs in only a few seconds for the same image. Table 1 shows the confusion matrix for the labeling on the same dataset.

A minor criticism we had of this dataset was its lack of diversity in scene categories. Out of the 715 images, more than 50% were street views, while about 10% were images of a animals in an open field (many of which were obtained from the MSRC dataset consisting mainly of a foreground object in the background). This dataset was more object-centric than general scene-based. **Note:**out of 715 images, there were only 14 images with the mountain label, hence its low diagonal score in Table 1.

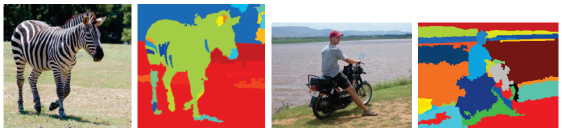


Figure 5: Examples of images and our segmentation output.

Human-in-the-loop Evaluation: To test our sentence generation paradigm, we compiled a new dataset consisting of 709 images. The dataset included many more different scene categories and objects than the Stanford dataset, although we "borrowed" several images from their dataset as well as from Hoeim et al., (Hoiem, Efros, and Hebert 2007). About 75% of the images in the dataset are taken in daylight while the remaining images are equally distributed over having sunset (or dusk), foggy (or snowy) and night-time illumination. The images with different scene illumination types were obtained from the Corel dataset as well as from *google images*.

Similarly, the distribution of scene categories is given as Mountains - less than 5%, Woods - less than 10%, Beach (or coastal) scenes - about 25%, Street - about 20%, Highway - about 5%, Suburbs - about 5%, In-city - about 15%, Garden - less than 10%, Tall buildings about 10%, and Open plains - less than 10%. It is important to note that many images strongly possess more than one scene category and are therefore counted multiple times. Some examples of images from our dataset are shown in Figure 1 and Table 2. We then used our *DISCO* paradigm, described in Sections 3 and 4 to generate sentences for each image.

Our human-in-the-loop evaluation was conducted as follows:

- The evaluation begins with the user accessing `http://www.zmldesign.com/SUnProject`. This webpage randomly selects one of the 709 images in the dataset and presents its generated text to the user.
- The corresponding image along with 19 others (a total of 20 images) are then presented to the user along with the generated text.

	Sky	**Tree**	**Road**	**Grass**	**Water**	**Buildings**	**Mountain**	**F_object**
Sky	**90.58**	1.37	0.40	0.00	0.09	6.09	0.14	1.32
Tree	1.41	**69.00**	0.62	2.07	0.02	22.15	0.14	4.58
Road	0.07	0.43	**87.58**	0.77	1.94	4.64	0.07	4.50
Grass	0.10	7.05	5.24	**80.81**	0.01	3.93	0.022	2.62
Water	4.74	0.33	24.34	0.18	**59.11**	4.63	1.28	5.39
Buildings	1.01	3.40	1.52	0.32	0.00	**85.42**	0.00	8.33
Mountain	6.93	18.99	10.65	4.91	6.08	37.13	**6.39**	8.90
F_object	1.03	3.97	8.76	0.54	0.23	21.78	0.00	**63.67**

Table 1: Confusion matrix (corresponding to our recorded highest accuracy) from labeling the Stanford dataset

- The user is encouraged to view all 20 images in detail and select the image most appropriately described by the given text.
- A survey page is launched and the user is encouraged to fill the questionnaire. The questions on the survey include: (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 people and color or shape in the presented sentences.

We conducted the web-based evaluation test with 25 users. Each user was instructed to test about 10 times and this was the case on average. We evaluated (i) the number of times the users correctly selected the same image as was being described; (ii) the user ratings of the description text compared to the image and (iii) the rankings of the order of usefulness of including different visual content information in the generated sentences.

Because multiple images can have very similar text descriptions, the user is asked to rate the appropriateness and meaningfulness of the sentence compared to expected selection. We provide this option so that even when 2 images correspond to similar sentences, we are still able to measure the meaningfulness of the sentences with respect to the expected image. The user is also presented with a set of questions to determine which image aspect: scene categorization, background description, object detection or object color and shape, yields the "best" information in the given set of sentences. Figure 6 shows the distribution of how users ranked the sentence descriptions provided. With total scores of *Fairly explains (3)* and higher, the algorithm was **60.34%** meaningful. With scores of *Satisfactorily explains (2)* and higher, the algorithm was 32.8% meaningful. **33%** of selected images were exactly matched to the expected image, although this is not conclusive evidence for sentence meaningfulness, since multiple images can have the same descriptions. Chance performance for *Fairly explains (3)* and higher would be 15%, for *Satisfactorily explains (2)* and higher would be 10%. If the test was run 250 times, chance performance of correctly selecting the right image each time would be less than 1%.

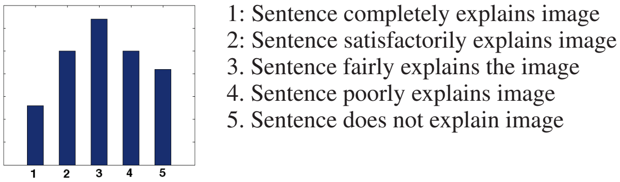


Figure 6: Distribution of meaningfulness scores of generated sentences

Table 2 shows sample results from the different types of image found in *DISCO*. Images tested include different scenes at sunset, at night-time, in foggy or snowy weather and in natural light. Several of the images shown are correctly labeled and have correct sentences, while others have incorrect labels, but with the sentences being quite general, are still descriptive and correct. Others yet have both incorrect labels and sentences, and the descriptions are misleading.

6 Discussion and Future Work

Sentences are extremely rich sources of information, both for transmitting and receiving information. In this paper, we have presented a sentence generation paradigm that is judged by its meaningfulness to humans and 60.34% of the tests performed indicated that the generated sentences at least fairly explained the presented images. Going forward, it will be useful to study what aspects of a scene yields the most information and under what circumstances. Our evaluation test website provides the opportunity for the further evaluation tests. 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, by encouraging our testers to provide more appropriate sentences for different images in the dataset, we are collecting human annotated data that will be useful going forward. We will also investigate how top-down models such as presented by (Farhadi et al. 2010) will integrate with a bottoms-up approach such as ours.

Acknowledgements

The authors would like to thank Randal Nelson for participating in the discussion leading to the development of the testing paradigm. We would also like to thank the anonymous reviewers for their insightful comments. IN is currently funded by the NSF Computing Innovations (CI) Fellowship (2010), under the NSF subaward No. CIF-384.

References

Alpert, S.; Galun, M.; Basri, R.; and Brandt, A. 2007. Image segmentation by probabilistic bottom-up aggregation and cue integration. In *CVPR*.

Comaniciu, D., and Meer, P. 2002. Mean shift: A robust approach toward feature space analysis. *TPAMI* 24(5):603–619.

Farhadi, A.; Hejrati, S. M. M.; Sadeghi, M. A.; Young, P.; Rashtchian, C.; Hockenmaier, J.; and Forsyth, D. A. 2010. Every picture tells a story: Generating sentences from images. In *ECCV*, 15–29.

Felzenszwalb, P. F., and Huttenlocher, D. P. 2004. Efficient graph-based image segmentation. *IJCV* 59:167–181.

Felzenszwalb, P. F.; Girshick, R. B.; McAllester, D.; and Ramanan, D. 2010. Object detection with discriminatively trained part-based models. *TPAMI* 32:1627–1645.

Gould, S.; Fulton, R.; and Koller, D. 2009. Decomposing a scene into geometric and semantically consistent regions. In *ICCV)*.

Gupta, A.; Srinivasan, P.; Shi, J.; and Davis, L. 2009. Understanding videos, constructing plots learning a visually grounded storyline model from annotated videos. 2012–2019.

Hoiem, D.; Efros, A. A.; and Hebert, M. 2007. Recovering Surface Layout from an Image. *IJCV* 75(1):151–172.

Li, L.-J.; Socher, R.; and Li, F.-F. 2009. Towards total scene understanding: Classification, annotation and segmentation in an automatic framework. In *CVPR*, 2036–2043.

Nock, R., and Nielsen, F. 2004. Statistical Region Merging. *TPAMI.* 26(11):1452–1458.

Oliva, A., and Torralba, A. 2001. Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope. *IJCV* 42(3):145–175.

Rabinovich, A.; Belongie, S.; Lange, T.; and Buhmann, J. M. 2006. Model Order Selection and Cue Combination for Image Segmentation. In *CVPR*, 1130–1137.

Reiter, E., and Dale, R. 2000. *Building Natural Language Generation Systems*. New York, NY, USA: Cambridge University Press.

Ren, X.; Fowlkes, C.; and Malik, J. 2005. Cue integration for figure/ground labeling. In *NIPS*.

Richard, X. H.; Zemel, R. S.; and perpi nán, M. A. C. 2004. Multiscale conditional random fields for image labeling. In *CVPR*, 695–702.

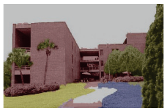
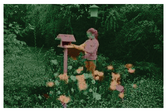
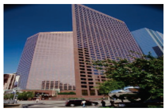
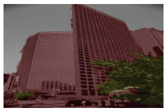
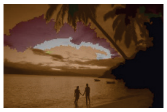
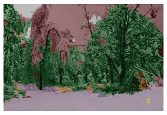
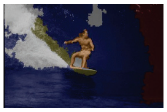
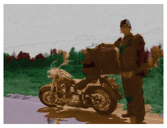
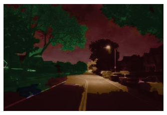
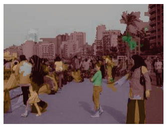
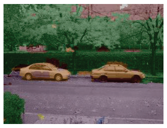
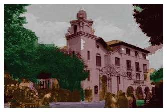
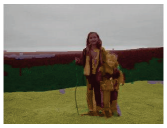
Shi, J., and Malik, J. 2000. Normalized cuts and image segmentation. *TPAMI.* 22(8):888–905.

Tu, Z.; Chen, X.; Yuille, A. L.; and Zhu, S.-C. 2005. Image parsing: Unifying segmentation, detection, and recognition. *IJCV* 63(2):113–140.

Tversky, B., and Hemenway, K. 1983. Categories of environmental scenes. *Cognitive Psychology* 49(15):121–149.

Xiao, J.; Hays, J.; Ehinger, K.; Oliva, A.; and Torralba, A. 2010. SUN Database: Large Scale Scene Recognition from Abbey to Zoo. In *CVPR*.

Yao, Z.; Yang, X.; Lin, L.; Lee, M. W.; and Zhu, S. 2010. I2t: Image parsing to text description. *Proc. of IEEE* 98(8):1485–1508.

Actual image	Regions labeled	Text generated
		The image scenery is set in the suburbs; There's quite a large expanse of white sky in the scenery. There's quite a large expanse of trees in the scenery. The wheat colored buildings are prominent in the scenery.
		The picture shows a scene... in the woods; The picture has a person in something white. The trees are quite prominent in the scenery.
		The picture shows a scene... with a view of tall buildings; The picture background consists mainly of bluish-purple sky. The picture background consists mainly of trees. The brown buildings are quite prominent in the scenery.
		This scene is at sunset or dusk... with a view of the water; There are at least 2 people in the picture.
		This scene shows a snowy day... in the woods;
		The image scenery is set with a view of the water; The picture has a person in something brown. The person is standing on the water. The green sky are quite prominent in the scenery. The picture background consists mainly of dark gray water. The dark gray buildings are quite prominent in the scenery.
		The picture shows a scene... with mountains in the background; The picture has a person in something dark gray. There is also a motorbike in the picture. There's quite a large expanse of dark gray trees in the scenery
		The picture shows a night scene... in a residential neighborhood outside the city;
		The picture shows a scene... 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.
		The image scenery is set in an open plains landscape; There are at least 2 cars in the picture including a purplish-pink one. The cars are on the road. There's quite a large expanse of green trees in the scenery.
		The image scenery is set 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 picture shows a scene... in an open plains landscape; There are at least 2 people in the picture. Someone in the picture has on something purplish-red. The people are standing on the sand. The dark gray buildings are quite prominent in the scenery. There's quite a large expanse of dark gray trees in the scenery.

■ sky □ cloud 🟪 road 🟩 foliage 🟦 water 🟥 bldg. 🟫 rock 🟧 fg obj. 🟨 sand

Table 2: Examples of images, labeled regions and generated sentences.