# Applications of Scene Attributes

Chen Xu
Computer Science Department
Brown University
chenx@cs.brown.edu

## Abstract

*In this paper, we study the feasibility of scene attributes as the intermediate scene representation for automatic image captioning, tag predicting and semantic image search. we show that when used as features for these tasks, low dimensional scene attributes can compete with or improve on the state of art performance. In particular, we propose a new method of content-based image retrieval, which takes advantage of the correlation between scene attributes and caption key words. When compared to simple uni-gram tf-idf image search, our method offers more promising results.*

## 1. Introduction

Patterson and Hays create and verify the SUN attribute database [7] in the spirit of analogous database creation efforts such as ImageNet [1], LabelMe [8] and TinyImages [9]. It is the first large-scale scene attribute database. In their work, they first derive a taxonomy of more than 100 scene attributes from crowd-sourced experiments. Next, they use crowd-sourcing to construct the attribute-labeled dataset on top of a significant subset of SUN database [11], spanning more than 700 categories and 14,000 images.

In order to use scene attributes for computer vision tasks, Patterson and Hays use the database to train classifiers for each attribute category to predict attributes. The final representation of scene attributes is a 102-dimensional feature representation, where each of the 102 values corresponds to the prediction confidence of a particular attribute in an image.

These simple attribute classifers have demonstrated their abilities to recognize a variety of attributes related to materials, surface properties, lighting, functions and affordances, and spatial envelope properties. In this paper, we want to investigate how this set of attribute classifers, as a low-dimensinal feature representation, can catch semantic information in images and improve the performance of tasks that require a deep understanding of image semantics. We carry out three experiments, including the areas of automatic image captioning, image tag predicting and content-based image retrieval.

## 2. Automatic Image Captioning

The im2text task [5] is to automatically generate a plausible caption for a given image. The published baseline first searches for the nearest neighbor of the query image according to some visual features, and then directly transfers the caption from the nearest neighbor to the input image. The authors investigate both gist descriptors and tiny images [9] as the features for global matching. Semantic similarity of captions is measured by BLEU [6] score. The experiments were carried out on the SBU Captioned Photo Dataset [1] which contains 1 million Flickr images with captions.

In im2text [5], the whole pipeline also includes a content matching step. Ordonez et al. rerank the nearest neighbors from global matching according to their content similarity to the query. The 5 kinds of image content are objects, stuff, people, scenes and the term frequency-inverse document frequency (TFIDF) weights. The authors then propose two ways, linear regression and linear SVM, to combine individual content measures into a final ranking. Their global plus content matching pipeline obtains BLEU scores 0.1215 +- 0.0071 for linear regression and 0.1259 +- 0.0060 for linear SVM on 1M dataset [5]. Note that these BLEU scores are not directly comparable to our performance scores in Table 1 because of some rounding in double precision of their released features.

The original im2text task uses captions that are not clean, which means the captions contain upper case letters, punctuations and stop words. The stemming process is not performed, either. The stemming process only keeps the root part of words, for example, stemming makes "run", "ran" and "runs" the same word "run". To better investigate how appropriate BLEU metric is for the image captioning task, we preprocess the captions. This includes removing punc-

[1] http://dsl1.cewit.stonybrook.edu/vicente/py/website/search

tuations and stop words, making all lower-case and stemming. In Table 1, the columns titled 10K*, 100K* and 1M* are the BLEU scores of the im2text task after preprocessing the 10K, 100K and 1M datasets. We observe that the scores are greatly decreased, and close to zero. We conclude that BLEU is not correlated well with semantic similarity between images and their captions.

Even though the scores are approaching zero, we find that the advantage of using scene attributes as features becomes much more significant. The BLEU scores obtained using scene attributes are about 1.4 times the BLEU scores obtained using baseline features(gist + tiny images). In 10K case, the chance score is close to the baseline results. This is further evidence showing the weakness of BLEU metric in captioning tasks.

It turns out that under the BLEU evaluation scheme used in im2text, much of the quantitative performance comes from chance matching of articles and prepositions between predicted and ground truth captions. As shown in Table 1, chance (random image retrieval) produces a score of 0.086 compared to 0.109 from the global feature baseline. We perform three operations to try and make the caption evaluation more rigorous: (1) stemming captions to root words, e.g. "run", "ran", "running" and "runs" are stemmed to "run", (2) converting all words to lower case and (3) removing frequent "stop words" such as articles and prepositions. While steps 1 and 2 make it easier for captions to match under the BLEU criteria, step 3 dramatically decreases performance as shown in Table 1, bottom. Chance performance drops by a factor of 6, to .014. The difference between attributes and the baseline global image features is more pronounced under this scheme – 0.0551 vs 0.0398, respectively. These numbers are quite low in absolute terms because the captions in the im2text database are exceedingly diverse, even for very similar scenes.

Fig. 1 shows some example results where scene attributes provide better global matching results on the im2text task than the results obtained from baseline features (gist + tiny images). For all examples, the generated captions obtained using attributes get higher BLEU scores than the generated captions from the baseline features. We observe that in all four cases, global matching using attributes returns images that are more semantically related to the query images. For example, in the first row in Fig. 1 the query is an image of grasslands, with a tree. Attribute based global matching returns the similar scene of grasslands, with a tree in the foreground. However, gist and tiny images based global matching returns an indoor scene. In the last row from Fig. 1, both query image and the image returned by attribute based global matching depict a furnished room, but the gist and tiny images based global matching returns a horse.

While our compact attribute representation improves data-driven image captioning over the global scene features, it does not represent state-of-the-art performance. Im2text [5] also investigates an image retrieval scheme which reranks the nearest neighbors based on recognized objects, materials, people, and scenes. This more sophisticated method outperforms our attribute-based retrieval, with BLEU scores up to 0.1259 +- 0.0060 on the unstemmed dataset.

Table 1: Global matching BLEU score comparison between baseline features and attributes on 10K, 100K and 1M dataset, 10K*, 100K* and 1M* are the dataset results with caption preprocessing(removing stop words, punctuations, stemming, all lower case)

| | 10K | 100K | 1M |
|---|---|---|---|
| Gist + Tiny Image | 0.0869 +- 0.002 | 0.0999 +- 0.009 | 0.1094 +- 0.0047 |
| Attributes | 0.0934 +- 0.01 | 0.1058 +- 0.015 | 0.1140 +- 0.0199 |
| Chance | 0.086 | | |

| | 10K* | 100K* | 1M* |
|---|---|---|---|
| Gist + Tiny Image | 0.02 +- 0.006 | 0.0255 +- 0.0079 | 0.0398 +- 0.0122 |
| Attributes | 0.0298 +- 0.0052 | 0.0366 +- 0.0132 | 0.0551 +- 0.0258 |
| Chance | 0.0144 | | |


Figure 1: *Attribute Search vs. Im2Text Baseline.* Example search queries that show how scene attributes provide more relevant image search results than the Im2Text baseline.

## 3. Predicting Image Tags

In this section, we investigate the performance of tag predicting tasks using 102 scene attributes as features. The dataset we use for these experiments is the MIRFlickr 25K dataset [4].

This set contains 25,000 images from the Flickr website, with user provided tags for each image. These Flickr tags are very noisy. Some tags may not be relevant to the image contents, such as the tag about camera model which was used to take the picture. Moreover, people tend to assign few tags to images instead of exhaustively listing all relevant tags.

More useful for our experiments are the 38 manual annotations of the dataset images. The images were annotated in two rounds. In the first round, 24 concepts are manually annotated and people are asked for each image whether it is at least partially relevant for each concept. In the second round, a stricter rule is used for 14 concepts. Only images labeled as relevant for a concept in the first round are considered here. They are considered relevant only if a significant portion of the image is relevant for the concept.

### 3.1. Tag Prediction with SVMs

We are very interested in how our scene attributes can do at tag predicting tasks. One of the state-of-art tag predicting methods is the TagProp model proposed by [2]. Later, Verbeek et al. applied the TagProp model to the MIRFlickr 25K dataset for tag predicting task and obtained promising results [10]. Verbeek et al. also compared their TagProp model to SVM classifiers for annotation predition [10], and found SVMs performed better than their model. They argued that by sacrificing the cost of precision, the training process of TagProp is faster than the training of SVMs.

Because [10] shows that SVMs for tag prediction always perform better than the TagProp model, we would like to train SVM classifiers to predict a confidence value for each of the 38 annotation categories using our scene attributes. We compare the performance of this method to the SVMs prediction results using the features proposed by the TagProp paper. According to [10], 15 distinct features are used, including one gist descriptor, 6 color histograms (RGB, LAB and HSV with 2 layouts), SIFT and hue descriptors (both with dense multi-scale grid and Harris-Laplacian detector, for 2 layouts). The TagProp features turn out to be 37,152 dimensions. In contrast, scene attributes provide a hugely more compact feature representation – only 102 dimensions.

Following the experiment setup of [10], we measure average precision (AP). To calculate the AP of a concept, we rank all images according to SVM confidence values and compute the precision at each position where the image is indeed relevant according to the manual annotation. The AP averages precision over all positions of relevant images. Table 2 shows the AP scores of 38 annotation concepts using scene attributes and 15 features, as well as the combination of all features. From our experimental results, scene attributes alone results in very promising prediction precision. The mean AP value over all annotation categories for scene attributes is 45.42%. This value is close to the results using the TagProp model. Verbeek et al. proposes two ways to define the training image weights, distance-based and rank-based weights, for the TapProp model, and the mean AP scores are 45.9% for distance-based weights and 46.5% for rank-based weights [10]. If we look at individual annotation categories, for some cases, like *structure* and *transportation*, the AP scores of attributes and 15 features are almost the same. We can say that with this compact feature representation, the 102 scene attributes still contains enough information for tag predicting tasks. An important thing to note is that most of the annotations are object-based, not attribute-based, like "baby", "car", etc. This may hurt the attributes' performance on this MIRFlickr dataset since there are no individual scene attribute classifiers for those particular objects. We also find that adding scene attributes and 15 features together does not improve AP scores significantly.

### 3.2. Principal Component Analysis Compression

We are impressed by the compactness and effectiveness of 102 scene attributes as features for tag predicting task. We want to further investigate the compactness of attributes. We decide to use principal component analysis (PCA) to reduce the dimensions of our feature space. We experiment using the first 32 principal components and first 16 principal components of the attribute features for the MIRFlickr images. Compared to using the full attribute features, performance decreases when fewer principal components are considered. when using 32 principal components, we get a mean average precision of 41.44%, which is acceptable. We get a mean average precision of 36.88% when considering 16 principal components.

We also quantize the 32 dimensional feature vector into binary values. The threshold between positive and negative labels is set at the mean of each dimension. This results in a 32-bit string feature representation. This feature representation obtain a mean AP score of 30.97%. Even with only 32 bits, attributes still do a much better job than chance.

### 3.3. Image Retrieval

**Attributes Words Correlation.** Our idea of applying attributes keywords correlations to image retrieval task is inspired by [3]. Mori et al. propose a method to count the number of co-occurrences of image patch based features and caption key words in a dataset of segmented images, to find correlations between key words and features. Mori el al. then use the calculated feature-keyword correspondence to predict the keywords in novel images. We discover the correspondence of attributes and keywords by counting the number of times that a given attribute and keyword appear in the same images. We also design a weighting scheme to make this method robust to noise. Our goal is to see if

Table 2: Comparison in terms of AP(%) of tag prediction with SVMs. 'Attr' is the results using 102 attributes as features, '15 feat' is the 15 features proposed by TagProp, and 'all' is Attr+15feat.

| | animals | baby | baby* | bird | bird* | car | car* | clouds | clouds* | dog | dog* | female | female* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Attr | 38.47 | 7.78 | 12.31 | 10.15 | 12.11 | 31.37 | 46.2 | 79.7 | 64.62 | 17.17 | 18.87 | 52.2 | 46.5 |
| 15feat | 48.01 | **12.39** | 16.18 | 16.40 | 22.57 | 34.68 | **51.57** | 84.84 | 77.12 | 29.97 | 32.94 | 58.42 | 54.75 |
| All | **48.09** | 12.25 | **16.19** | **16.47** | **22.68** | **34.88** | 51.51 | **84.96** | **77.20** | **30.16** | **33.10** | **58.46** | **54.78** |
| Rand | 12.9 | 1.1 | 0.5 | 3.0 | 2.0 | 5.0 | 1.7 | 14.5 | 5.4 | 2.7 | 2.3 | 24.8 | 15.9 |

| | flower | flower* | food | indoor | lake | male | male* | night | night* | people | people* | plant | portrait |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Attr | 39.46 | 44.36 | 40.01 | 71.37 | 24.1 | 47.46 | 37.18 | 60.73 | 48.54 | 76.45 | 69.81 | 74.89 | 59.53 |
| 15feat | 52.09 | 62.55 | 50.22 | 75.55 | **27.73** | 51.79 | 42.73 | **65.07** | 53.93 | 80.31 | 75.9 | 79.75 | 68.81 |
| All | **52.16** | **62.66** | **50.25** | **75.74** | 27.69 | **51.86** | **42.81** | 65.03 | **54.03** | **80.37** | **76.1** | **79.82** | **68.91** |
| Rand | 7.4 | 4.4 | 4.0 | 33.5 | 3.0 | 23.9 | 14.2 | 10.3 | 2.5 | 41.3 | 31.1 | 34.8 | 15.6 |

| | portrait* | river | river* | sea | sea* | sky | struct. | sunset | transp. | tree | tree* | water | Mean |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Attr | 59.26 | 21.83 | 6.73 | 51.51 | 28.86 | 84.26 | 78.43 | 58.22 | 44.06 | 61.85 | 40.86 | 58.57 | 45.42 |
| 15feat | 68.71 | **25.51** | **6.92** | **56.2** | 31.46 | 88.91 | 78.38 | 67.83 | 45.74 | **68.03** | 52.63 | 62.58 | 52.08 |
| All | **68.81** | 25.47 | 6.72 | **56.2** | **31.6** | **88.93** | **78.63** | **67.84** | **45.79** | 68.02 | **52.71** | **62.72** | **52.15** |
| Rand | 15.1 | 3.7 | 0.6 | 5.3 | 0.8 | 31.0 | 40.4 | 8.4 | 11.9 | 18.3 | 2.7 | 13.1 | 12.3 |

scene attributes convey semantic information and if they are suitable for image captioning or tag prediction tasks.

We use the 10,000 images and captions from im2text dataset as our training set. We only consider the 1000 most common words in the im2text dataset as key words. We let $n$ be the size of image dataset, and create an $n$-long vector $W_i$, for each word $w_i$ and an $n$-long vector $A_j$ for each attribute $a_j$. The $k$th element of $W_i$ indicates if the word $w_i$ exists in the caption of the $k$th example in the dataset. Similarly the $k$th element of $A_j$ indicates if the attribute $a_i$ exists in the image of the $k$th example in the dataset.

We also use a binary-idf, binary-inverse document frequency, style weighting for word vectors and tf-idf, term frequency-inverse document frequency, style weighting for attribute vectors. In detail, if $w_i$ exists in the caption of the $k$th example, the weight of the $k$th element in $W_i$ is set to be $1/f_w$, where $f_w$ is the inverse document frequency of $w_i$; otherwise the weight is zero. Similarly, if $a_j$ exists in the image of the $k$th example, the weight of the $k$th element in $A_j$ is set to be $conf/f_a$, where $conf$ is the sigmoid scene attribute confidence score, and $f_a$ is the inverse document frequency of $a_j$; otherwise the weight is zero. Finally, the correlation between word $w_i$ and attribute $a_j$ is simply the inner product of $W_i$ and $A_j$ - $C_{ij} = W_i * A_j$.

Table 3 shows top correlated words for attributes and Table 4 shows top correlated attributes for words. We set a threshold -0.75 on the SVM confidence to determine if a particular attribute exists in the image. We find that attributes and keywords are semantically correlated. Looking at the top correlated key words for attribute 'sailing/boating', the top correlated key words are 'cruise', 'harbor', 'ocean', 'sail', 'swim', etc. Note some words are transformed because of stemming. Intuitively these key words are related to 'sailing' and 'boating'.

**Word-to-Attribute Correlation Applied to Image Retrieval.** In this section, we apply the word-to-attribute correlation scores to the image retrieval task. We want to achieve content-based image retrieval with text queries. For example, if user inputs a text query "sky", we can convert the text to its corresponding visual features, such as blue background, clustered clouds and horizon line, and retrieve images using those visual features. This way we do not have to look at the image captions in the database. Here we use the attribute-keyword correlation we have obtained to do the conversion from text features to visual features. We again use the im2text dataset, with 10,000 examples for training, and 90,000 examples for testing.

Table 3: Examples of top correlated words for attributes. Note: words are stemmed.

| Attr. | **sail./boat.** | **driving** | **eating** | **railroad** | **camping** |
|---|---|---|---|---|---|
| Top 20 Correlated Words | cruis | sand | bar | moon | grass |
| | harbor | road | cabinet | railwai | pastur |
| | ocean | sidewalk | desk | lit | field |
| | sail | lane | kitchen | exposur | forest |
| | swim | dune | oven | harbour | landscap |
| | boat | highwai | tv | track | fallen |
| | dock | moon | een | southern | lone |
| | sunset | traffic | shelf | train | hidden |
| | sky | canyon | breakfast | mother | hill |
| | airplan | track | dine | star | flow |
| | beach | wind | tabl | light | stream |
| | sea | order | ceil | tank | canyon |
| | coast | cross | candl | traffic | oak |
| | wave | bridg | lit | night | trail |
| | ski | cabl | sunris | glow | distanc |
| | clear | ga | chocol | pass | road |
| | lake | drive | second | shadow | camp |
| | ship | fallen | room | salt | creek |
| | moon | colorado | bathroom | site | grow |
| | sunris | toward | cherri | wing | wind |

Given the query text, we break the text into words. Letting $T_{query}$ be the vector of query word indices. These indices are the positions of the query words in the list of 1000 most common caption words. We use $T_{query}$ and word-attribute correlation we have obtained to create an estimated scene attributes representation. We call these estimated attributes 'fake' attributes in this paper. Each word $w_i$ has a vector of correlations $C_i =< c_{i,1}, ..., c_{i,j}, ..., c_{i,102} >$,

Table 4: Examples of top correlated attributes for words

| Words | **kitchen** | **mountain** |
|---|---|---|
| Top 10 Correlated Attr. | tiles | far-away horizon |
| | enclosed area | hiking |
| | cleaning | camping |
| | reading | natural |
| | wood (not part of a tree) | foliage |
| | glossy | vegetation |
| | electric/indoor lighting | trees |
| | glass | rugged scene |
| | eating | shrubbery |
| | studying/learning | leaves |

| Words | **beach** | **dress** |
|---|---|---|
| Top 10 Correlated Attr. | ocean | cloth |
| | far-away horizon | medical activity |
| | sand | enclosed area |
| | waves/surf | paper |
| | sunbathing | no horizon |
| | sailing/boating | sterile |
| | diving | research |
| | swimming | electric/indoor lighting |
| | still water | stressful |
| | open area | man-made |

where each element $c_{i,j}$ is the correlation of word $w_i$ and attribute $a_j$. The fake scene attribute representation is defined as the average of correlation vectors of the words in the query,

$$F_{fake} = \frac{1}{N}\sum_{k=1}^{N} C_{T_{query,k}} \quad (1)$$

where N is the length of $T_{query}$, and $T_{query,k}$ is the $k$th element of $T_{query}$, the index of the $k$th query word in common words list. We consider the same word multiple times if it appears multiple times in the caption.

We then learn multi-linear regressions to map fake scene attributes to predicted scene attributes, which are the output feature vectors of attributes classifiers. In the training dataset, for each image, we know both its fake attributes and predicted attributes. We then learn the regression to map from those attributes in the fake representation to $a_j$ in the predicted representation. Finally, we search for the nearest neighbors of the query's predicted attribute representation in the test dataset.

We compare our method to tf-idf based image retrieval method because tf-idf is a widely used baseline method for text-based image retrieval. Fig. 2 and 3 show the results of both methods separately. From the results, we can see that attribute-based image retrieval gives very promising search results. For most search results returned by attribute-based method, the target specified by the query text are the dominant objects or scenes in the retrieved images. However, that is not the case for tf-idf based method. For example, for the "flower" query, the five images returned by attribute-based method are all depicting flowers directly, while the dominant objects in images returned by tf-idf based method contain mug, pony and bee. Another interesting thing we find is that our method can understand the semantics of not only single words but also phrases. For example "snow mountain" and "dark sky", most of the results have the correct semantics. For the "dark sky" example, there are some false positives, such as the second image where a boat is agaist the blue sky. It seems our method thinks the relatively dark boat is a part of the sky.

## 4. Discussion

We have seen that using scene attributes as features can compete with or improve on the performance of several computer vision tasks, including automatic image captioning, tag predicting and content-based image retrieval. However, the applications are not limited to these areas. There are other computer vision tasks requiring features providing sufficient semantic information, such as automatic image illustration. In our future experiments, we will further explore the interplay between scene attributes and image semantics through a variety vision tasks.

Currently, our proposed image retrieval method enables us to input keywords or phrases for image search. However, our goal is to make the algorithm understand more complicated input queries, such as a whole sentence. For image captioning, we also need to find a better metric to measure the semantic similarity between captions and images, since BLEU does not work well.

## References


[1] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009.

[2] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid. Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In *International Conference on Computer Vision*, pages 309–316, sep 2009.

[3] Y. M. Hironobu, H. Takahashi, and R. Oka. Image-to-word transformation based on dividing and vector quantizing images with words. In *in Boltzmann machines, Neural Networks*, page 405409, 1999.

[4] M. J. Huiskes and M. S. Lew. The mir flickr retrieval evaluation. In *MIR '08: Proceedings of the 2008 ACM International Conference on Multimedia Information Retrieval*, New York, NY, USA, 2008. ACM.

[5] V. Ordonez, G. Kulkarni, and T. L. Berg. Im2text: Describing images using 1 million captioned photographs. In *Neural Information Processing Systems (NIPS)*, 2011.

[6] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

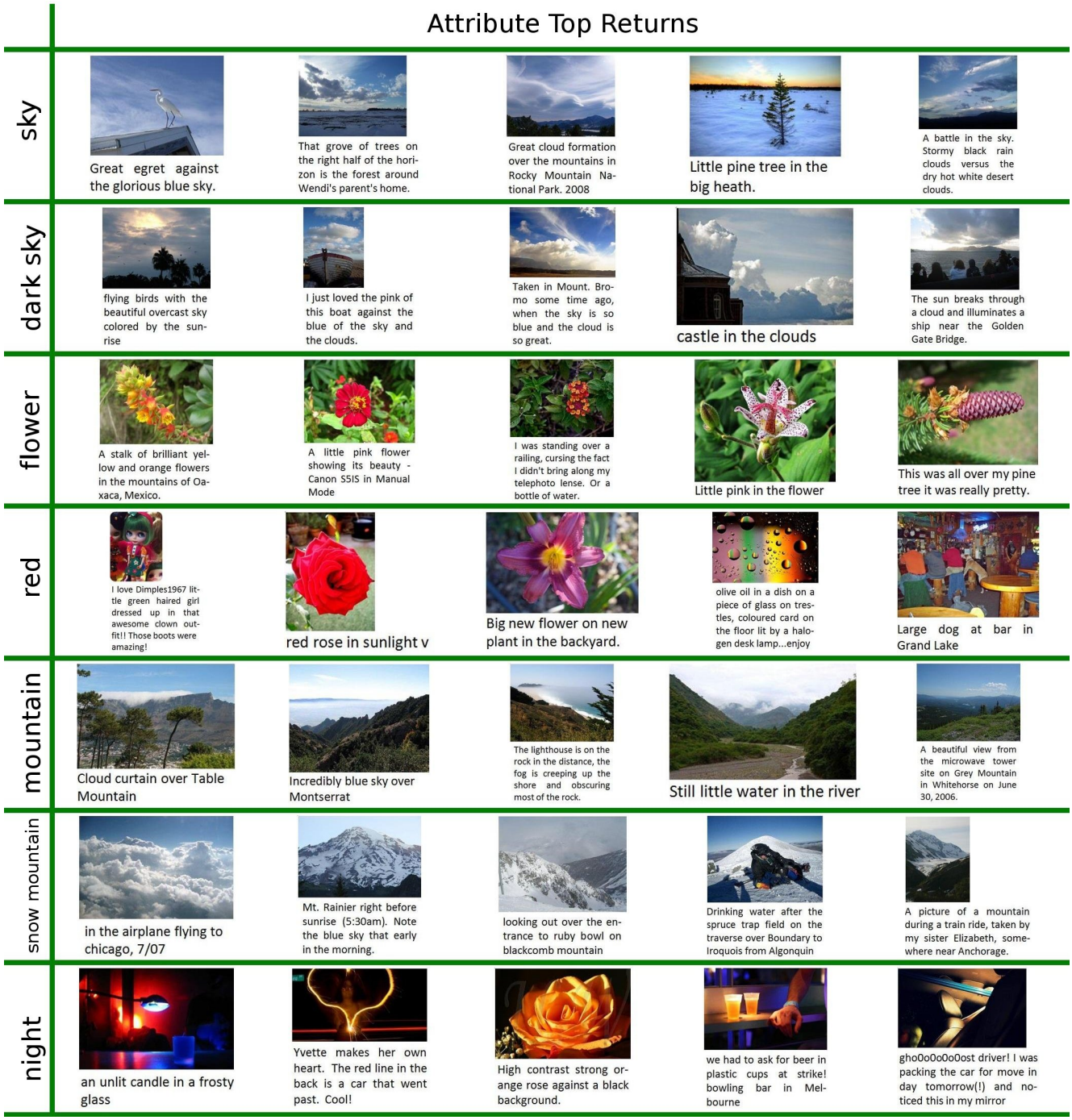


Figure 2: *Attribute Based Image Retrieval Results on 1M Im2text Dataset.*Although the captions are shown here for completeness, our text-based image retrieval method did not see them at query time, whereas the TF-IDF method (Figure 3) uses the captions exclusively.


[7] G. Patterson and J. Hays. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *Proceeding of the 25th Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[8] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. LabelMe: a Database and Web-based Tool for Image Annotation. *IJCV*, 77(1-3), 2008.

[9] A. Torralba, R. Fergus, and W. T. Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(11):1958–1970, Nov. 2008.

[10] J. Verbeek, M. Guillaumin, T. Mensink, and C. Schmid. Im-

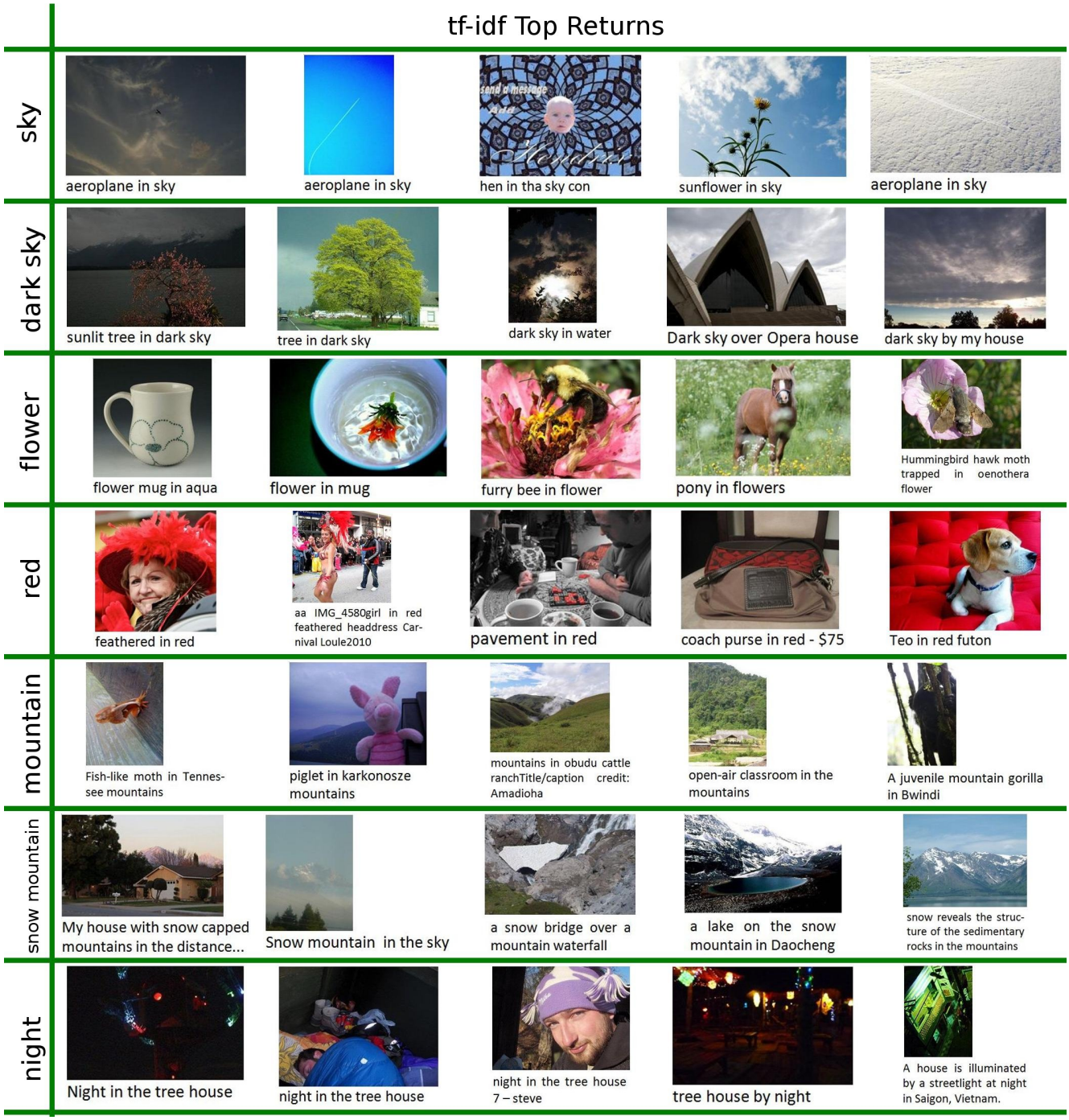


Figure 3: Tf-idf Based Image Retrieval Results on 1M Im2text Dataset

age Annotation with TagProp on the MIRFLICKR set. In *11th ACM International Conference on Multimedia Information Retrieval (MIR '10)*, pages 537–546, Philadephia, United States, 2010. ACM Press.

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