Large Image Collection Visualization Using Perception-Based Similarity with Color Features

Zeyuan Chen and Christopher G. Healey

Department of Computer Science, North Carolina State University, USA

Abstract. This paper introduces the basic steps to build a similarity-based visualization tool for large image collections. We build the similarity metrics based on human perception. Psychophysical experiments have shown that human observers can recognize the gist of scenes within 100 milliseconds (msec) by comprehending the global properties of an image. Color also plays an important role in human rapid scene recognition. However, previous works often neglect color features. We propose new scene descriptors that preserve the information from coherent color regions, as well as the spatial layouts of scenes. Experiments show that our descriptors outperform existing state-of-the-art approaches. Given the similarity metrics, a hierarchical structure of an image collection can be built in a top-down manner. Representative images are chosen for image clusters and visualized using a force-directed graph.

1 Introduction

With the rapid development of cellphone cameras, storage devices and social media, capturing, saving and sharing images has become much more common. Because of this, the size of image collections is growing rapidly both in terms of absolute data size and the number of images being saved. For example, Flickr is reported to contain 5.26 billion public photos, with approximately 2 million photos uploaded each day in 2015 [1]. Instagram reports 95 million photos are being posted daily [2]. Even a personal gallery may contain many thousands of images. This makes it difficult or impossible for users to retrieve images or fully explore a large image collection by visual inspection. Therefore, we seek an effective exploration tool that allows users to interactively and efficiently browse and retrieve images from large collections in ways that keep the images organized. Although images are generally classified as object images or scene images, in this paper we focus on scene images.

Visualization is an effective way to communicate information through visual imagery. Currently, there are few examples of visualization tools for large image collections. Most exploration tools manage images using labels like dates, events, people and locations. Unfortunately, this method has numerous drawbacks. First, users usually do not provide detailed labels, which makes it difficult to retrieve images by keyword query when they are incomplete or unavailable. Second, many words are ambiguous. For example, if a user searches on the label *apple*, it could be a fruit or a personal computer.

Recent works [3–5] have used similarity-based approaches to visualize large image collections. The images are grouped into clusters automatically based on their visual distances. There are two main issues: developing the metric of similarity between images, and visualizing clusters. First, the computational distances between images should be similar to their perceived difference. Many traditional computer vision approaches [6, 7] describe images based on lower-level features (e.g. textures and edge-based structures). This can make them slow or inaccurate. We want to design similarity metrics by learning from human perception. Humans can understand the gist of a scene in a single glance, even if the image is blurred, which means people often recognize the categories of images as a whole without looking at significant amounts of low-level detail [8]. Recent works based on global features like the "gist" of Oliva and Torralba [9] and the spatial pyramid framework of Lazebnik et al. [10] have achieved great performance both in terms of accuracy and efficiency. Second, it is infeasible to display all the images in clusters at once, so it is crucial to find appropriate ways to visualize clusters.

In this paper, we present the following novel contributions: (1) a method to process scene descriptors in a way that preserves both color information and global scene properties; (2) an approach to construct highly compact scene descriptors; (3) an application of the Color Coherent Vector (CCV) model [11] to preserve coherent color region information, improving the performance of the scene descriptors; (4) and classification experiments on both small and large image datasets to statistically verify that our methods outperform existing state-of-the-art approaches.

2 Background

Human observers can understand a variety of visual information from an image and recognize its basic-level category within 100 msec [12]. Instead of relying entirely on computer vision techniques, there are potential advantages in learning from human observers' perceptual processing of images.

2.1 Human Perception of Scene Gist

The experimental work of Greene and Oliva [13] has shown that in the early perception stages, observers can comprehend global properties of a scene (such as the mean depth of the scene and whether it is navigable or not) more easily than classifying them into basic-level categories (e.g. mountains, rivers). This result supports the user study [14] in which seven global properties were chosen to represent scenes: openness, expansion, mean depth, temperature, transience, concealment and navigability. In the fast scene categorization task, the performance of human observers was indistinguishable from a Naive Bayes classifier whose input was rankings for the seven properties. These results suggest global properties are sufficient to represent the gist of a scene.

The role of color in rapid recognition of scene gist is significant. The experiments of Oliva and Schyns [12] have shown that color has an influence on recognition tasks when the color is related to the meaning or "diagnostic" of the scene categories. The experimental study of Castelhano and Henderson [15] demonstrated that gist activation is affected by color when the images are blurred. The study also showed the reason color expedites gist activation is not because it helps in the segmentation process, but because it is directly related to the scene's gist.

2.2 Scene-centered Image Features

Many traditional scene recognition approaches are object-centered. They build hierarchical levels of features in a bottom-up manner [6, 7]. The lower-level features (e.g. colors and textures) are grouped into higher-level ones (e.g. regions and objects).The semantic meanings of the scenes are inferred from the highest-level features.

However, recent scene-centered approaches [16, 9, 17] have shown that holistic representations of images can be built directly from low-level features. This is because there are regular and unique patterns of statistical distributions of features in different scene categories. The global properties of a "gist" descriptor [9] are estimated from a set of global feature templates that are effective to represent images.

Different from approaches based on the human perception of scene gist, related research [18, 10, 19] makes use of state-of-the-art computer vision techniques to build global features. Xiao et.al [18] identify several features as kernels including histogram of oriented gradients, scale-invariant feature transform (SIFT), color histograms, etc. The "all features" classifier that is based on a weighted sum of these kernels is reported to have higher precision than any individual feature during scene classification. Alternatively, the spatial pyramid framework [10] is a sophisticated method to summarize local features. Images are iteratively partitioned into sub-regions, and histograms of features are computed locally in each sub-region. Similar to the "gist" descriptor, this framework preserves the spatial layouts of features and is widely used in scene classification.

2.3 Color-based Image Features

Color information plays an important role in scene recognition. A color histogram itself is not enough for complex recognition tasks, but color histograms with geometric information have achieved strong performance [20, 11, 21]. The CCV model [11] computes histograms for coherent and incoherent pixels separately. Pixels in large contiguous regions are considered coherent while the remaining pixels are incoherent.

2.4 Image Collection Visualization

After the image features are extracted, pairwise distances between images can be computed. The images are generally grouped into clusters using standard methods like k-means based on their pairwise distances. For example, Google

Image Swirl [5] builds clusters of images based on visual similarity. An exemplar is chosen from each cluster as a representative. The exemplars are visualized using a balloon tree.

3 Similarity Metrics

There are two important findings we can harness from human perception. First, global properties are sufficient for scene recognition and scene descriptors can be built directly from low-level features. Second, color information helps scene recognition. State-of-the-art descriptors like "gist" [9] and histogram of oriented gradients with a spatial pyramid framework (denoted HOG) [10] have preserved the global features, but they only deal with grayscale images. There are many suggestions for color descriptors, such as RGB-SIFT [22] and HSV-SIFT [23]. They compute SIFT descriptors in different color channels separately, then stack the feature vectors together. However, the size of those descriptors are several times larger than the original ones, which makes them less efficient. In this paper, we will present novel approaches to build compact descriptors that preserve both color information and global properties.

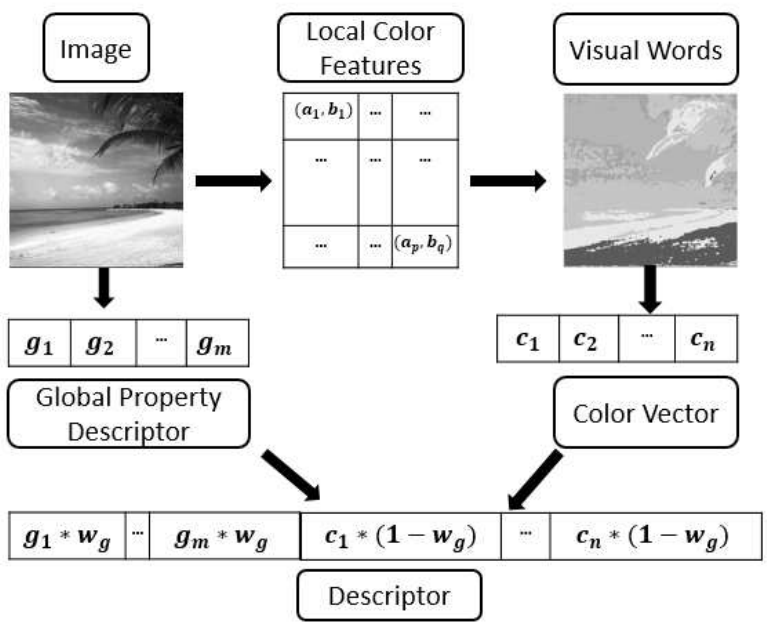

Fig. 1. Pipeline of our approach. (a_i, b_i) is the color feature for each pixel. $(g_1, ..., g_m)$ is a global property descriptor. $(c_1, ..., c_n)$ is a color vector. w_g is the weight of global property descriptor in feature combination

The pipeline of our approach is shown in Figure 1. The local color features are extracted densely and then are quantized into "visual words" using the bag-of-words model. An image is represented by a color vector computed from the frequencies of "visual words" in the spatial pyramid framework. The final descriptor of an image is the weighted combination of the color vector and a global property descriptor (e.g. "gist" or HOG). The details are discussed in the sections below.

3.1 Color Feature Extraction

In the experiments of Oliva and Schyns [24], the (a, b) vector of Lab color space is used to represent colors based on two important Lab properties. First, the luminance L is separated from the other two channels, so (a, b) is invariant to changes of luminance in the scene. Second, a and b represent colors along two color-opponent dimensions. The human visual system processes colors along similar dimensions. Additionally, the Euclidian distance between two (L, a, b) vectors is similar to their perceived color difference.

After converting an image from RGB space to Lab, each pixel is described by a (L, a, b) vector. We discard L for lightness change invariance and only use the (a, b) vector as the color feature for each pixel.

3.2 Feature Quantization and Image Representation

The local color features are quantized to n discrete colors using the bag-of-words models described in [25]. The local features are randomly sampled from the image collection and are then grouped into n clusters using k-means. The n cluster centers are called the "vocabulary". The "visual word", or the quantized values, of each pixel in the image is the index of its nearest center.

The CCV model [11] shows that splitting an image into coherent and incoherent regions contributes to the improvement of image retrieval accuracy, which matches the experimental results of Castelhano et al. [15] that shows color regions are related to human scene recognition. A pixel is coherent if it is in a color region whose size is greater than a threshold T. Given the total number of pixels in an image N, the normalized threshold t is given by $t = \frac{T}{N}$.

In order to differentiate coherent and incoherent pixels in our approach, we update the visual word of each incoherent pixel using

$$VW_i = VW_i + n \tag{1}$$

Here, VW_i is the visual word (which is a scalar value) of the i-th pixel in an image and n is the length of the "vocabulary".

The spatial pyramid framework [10] is used to summarize the visual words. This framework divides images into sub-blocks at different levels and computes histograms within each sub-block. The histograms are stacked together with weights to generate a vector to describe the image. By applying Equation 1, it is equivalent to stacking histograms of coherent and incoherent pixels together in each sub-block. Therefore, the coherent and incoherent information does not break the overall spatial layout of the spatial pyramid framework.

3.3 Feature Combination

In scene classification, the weighted combination of multiple kernels can generate better results if the weights are chosen properly. The weights of individual kernels of the "all features" descriptor of Xiao et al. [18] are proportional to the fourth power of their accuracy. Hou et al. [26] assigns higher weights to kernels with high accuracy. Past research has shown color features alone cannot achieve high accuracy. That is probably why color information is often neglected. However, the study of human perception [12, 15] shows the importance of color in scene recognition. Instead of combining a large number of different features, we focus on combining color and global features.

The kernel combination method is designed for supervised learning, but we want to apply our approach to unsupervised clustering of images so that the image clusters can be constructed more flexibly. Let G be a global feature descriptor and C the computed color vector described in above sections. The descriptor D of an image is the weighted concatenation of G and C as shown in this equation:

$$D = CAT(w_g * G,\ (1 - w_g) * C). \tag{2}$$

Here, w_g is the weight of the global-property descriptor. Suppose X and Y are two sets of values (e.g. a collection of visual words) and H_X and H_Y are their histograms that both have m bins. The histogram intersection kernel between H_X and H_Y is given by the *histogram intersection function* [27]:

$$K(H_X, H_Y) = \sum_{i=1}^{m} min(H_X(i), H_Y(i)). \tag{3}$$

Suppose G is a descriptor built from histograms (e.g. HOG), and $K(G_X, G_Y)$ is a kernel of feature G. Let $K(C_X, C_Y)$ be a kernel of C, K_{com} be the weighted combination of the kernel of G (weighted by w_g) and C (weighted by w_c), and K_D be a kernel of D. Then K_{com} is:

$$K_{com} = w_g * K(G_X, G_Y) + w_c * K(C_X, C_Y). \tag{4}$$

If w_c = (1-w_g), K_D is:

$$K_D = K(CAT(w_g * G_X,\ w_c * C_X),\ CAT(w_g * G_Y,\ w_c * C_Y)). \tag{5}$$

We can prove that K_{com} equals K_D given w_c = (1-w_g), which illustrates that constructing the combined feature D using Equation 2 is equivalent to weighted kernel combination if G is in the form of a histogram. A global feature descriptor like "gist" is not built directly from histograms, but it also summarizes low level features. Therefore, Equation 2 works for other global feature descriptors like "gist" as well.

The descriptor D is constructed in a very compact way since the vocabulary of C is usually small. For instance, for a two-level HOG descriptor with a 200-word vocabulary and two-level color vector C with an 8-word vocabulary, the length of the HOG descriptor is 1000 while the length of the D vector is only 1080.

4 Experiments

We use scene classification experiments to evaluate the performance of our descriptors. Since the categories of images in the database are labeled by observers, the classification accuracy measures how different the similarity metrics are from human-perceived visual distances. The first part of the experiment is to find optimal values for w_g (the weight of the global feature descriptor, denoted w for simplicity) and t (the normalized threshold to split coherent and incoherent pixels). The second goal is to compare our algorithms with other state-of-the-art approaches to scene classification. The "gist" [9] (denoted *gist*) and HOG [10] approaches are chosen for comparison.

There are two datasets used in the experiments. The first is the Eight Scene Categories (8-scene) Dataset [9] with 2,688 color images from eight outdoor scenes. The second is the SUN397 [18] Dataset with 108,604 color images from 397 categories including indoor, outdoor natural and outdoor man-made scenes. There are at least 100 images in each scene category in both datasets. The 1-vs-all SVMs were trained using samples from those two datasets.

As described above, the descriptor D could be constructed in multiple ways. First, the CCV model could be applied (denoted *ccv*) or not (denoted *noccv*). Second, the color vector C could be combined with different global feature descriptors. There are four possible types of descriptors: *gist-ccv*, *gist-noccv*, HOG-*ccv*, and HOG-*noccv*.

4.1 Parameter Selection

The optimal parameters of the above four descriptors were computed from small sampled image sets of a database. For the 8-scene database, 50 training and 50 test images were randomly chosen from each category for each trial. There were 20 trials in total. The sample size was about 30% of the database. For SUN397, 10% of the images were uniformly selected to construct an image subset, then five training and five test images were randomly chosen for each category from the subset for each trial. Five trials were run. The size of samples for each trial was 3.65% of SUN397 database.

When the CCV model was not applied, w was the only parameter. There were two parameters w and t when CCV was applied. As an example, in the 8-scene database, the classification accuracy of HOG-*ccv* and HOG-*noccv* with changes of w and t is shown Figure 2.

We simply chose the w and t associated with the highest average accuracy as their optimal values (shown in Table 1). In each case, the "slope" of the accuracy

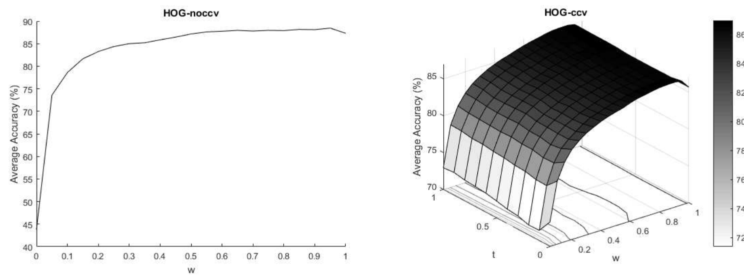

Fig. 2. The average accuracy of different descriptors with w and t changing from zero to one. Left: results of HOG-*noccv* descriptors. Right: results of HOG-*ccv*

Table 1. The optimal parameters for all descriptor in the 8-scene and SUN397 database

	gist-noccv	*gist-ccv*	HOG-*noccv*	HOG-*ccv*
8-scene	w = 0.70	w = 0.55, t = 0.95	w= 0.95	w = 0.90, t =0.05
SUN397	w = 0.75	w= 0.70, t = 0.95	w = 0.80	w = 0.70, t = 0.95

was small near the optimal values. This suggested the accuracy of descriptors was not sensitive to changes in w or t.

4.2 Comparison with State-of-the-Art

The four combined descriptors were compared with *gist* and HOG respectively. Given the optimal values of w and t, the experiments were run on the 8-scene database (50 trials) and the SUN397 database (10 trials). In each trial, 50 training and 50 test images were randomly selected from each category. All of the descriptors used the same training and test images.

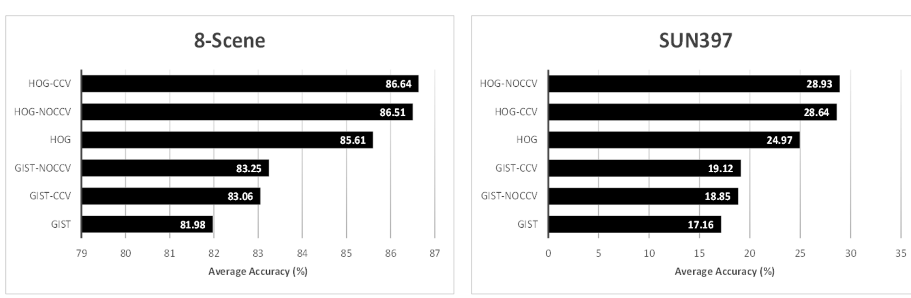

Fig. 3. The classification accuracy of different descriptors in 8-Scene and SUN 397 database

The average classification performances are shown in Figure 3. For each database, we performed two groups of repeated measures one-way analysis of variance (ANOVA) and post-hoc tests to compare the accuracy of G (HOG in group one and *gist* in group two) with their combined descriptors D (shown in Table 2, including F-values). In both 8-scene and SUN397 database, the combined descriptors D were all significantly more accurate than the original global feature descriptors G alone. Additionally, there were no significant differences between *G-ccv* and *G-noccv* in the 8-scene database, but *gist-ccv* was significantly more accurate than the *gist-noccv* while the accuracy of HOG-*noccv* was significantly higher than HOG-*ccv* in SUN397. This suggested that the CCV model worked well with gist-based descriptors (e.g. the "gist" [9]), but not so well with computer-vision-based descriptors (e.g. HOG [10]) in a large database. Moreover, we found that *gist-ccv* performed consistently better than *gist-noccv* for every trial in SUN397, which indicated that the color coherent information had a significant influence on the improvement of classification accuracy when combined with gist-based descriptors.

Table 2. The pairwise comparisons among the accuracy of state-of-the-art descriptors (i.e.*gist* and HOG, denoted G) and the combined descriptors. The first two rows are the results on the 8-Scene database and the last two rows are on the SUN397 database

Group	G – *G-ccv*	G – *G-noccv*	*G-ccv* – *G-noccv*	F
HOG_8	-1.030, $p<0.01$	-0.900, $p<0.01$	0.130, $p=0.702$	$F(2,48)=38.449, p<0.01$
$gist_8$	-1.080, $p<0.01$	-1.275, $p<0.01$	-0.195, $p=0.403$	$F(2,48)=41.004, p<0.01$
HOG_{397}	-3.669, $p<0.01$	-3.962, $p<0.01$	-0.293, $p<0.01$	$F(2,8)=1852.262, p<0.01$
$gist_{397}$	-1.967, $p<0.01$	-1.698, $p<0.01$	0.269, $p<0.01$	$F(2,8)=435.452, p<0.01$

5 Cluster Visualization

After the descriptors are computed for every images in a dataset, the images can be grouped into clusters using k-means. The visual distance between two images is defined by the distance of their descriptors. A tree structure of clusters is built in a top-down and breadth-first manner. Let L_i be a set of clusters at the i-th level. Starting from the entire dataset (L_0), the dataset is divided into k clusters (L_1). Every cluster in L_i is split to construct L_{i+1}. A cluster cannot be split once its size is smaller than a certain threshold. The process terminates when there are no more clusters to split. The number k of k-means clustering is determined using the gap statistic method [28] or set by a user.

One image is chosen from each cluster to represent it. The selected image should be as similar to all other images as possible. The distances of every pair of images are computed, and the image with the smallest average distance from all other images is chosen as the representative.

Fig. 4. The visualization of clusters at level one. The Dataset is the 8-scene database

To achieve intuitive and aesthetic visualization results, the distances of images should be preserved and occlusions should be avoided if possible. To achieve this, a force-directed graph is employed. The visualization of clusters at level one of the 8-scene database is shown in Figure 4. The on-screen Euclidean distance represents the visual differences between images: farther apart for larger dissimilarities. The overall distribution of images is uniform and relatively symmetric.

If the users want to see the details of a cluster, they can expand (or "zoom into") it. Then, only the representative images of the sub-clusters of the expanded cluster will be shown. A slider bar on the top of the screen shows the current level of the clusters being displayed.

6 Conclusion

We have introduced the basic steps for constructing a similarity-based large image collection visualization system based on gist-based similarity metrics. Inspired by the exceptional ability of humans to perceive scenes, we designed global feature descriptors that exploit color informaiton. Our descriptors preserve the information from coherent color regions, as well as the spatial layouts of scenes. The color features are combined with sophisticated global property features in a compact way. Parameter selection experiments were run on small sampled images sets to find optimal parameters for our descriptors. The experimental results showed that our descriptors are insensitive to changes in those parameters. Follow-on similarity experiments identified our approaches as achieving significantly improved precision over state-of-the-art algorithms like "gist" [9] and HOG with spatial pyramid framework [10].

We have also introduced methods to build hierarchical structures for image collections given our similarity metrics, and have shown how to use these hierarchies to visualize image clusters using force-directed graphs.

We note that a single image may not be able to represent clusters with large variance. It is better to synthesize representative images that are in accordance with the users' impressions of clusters. This is left for future work.

References

1. Michel, F.: How many public photos are uploaded to flickr every day, month, year? (2015) Accessed: 2016-8-5.
2. https://www.instagram.com/press/ (2016) Accessed: 2016-8-5.
3. Nguyen, G.P., Worring, M.: Interactive access to large image collections using similarity-based visualization. Journal of Visual Languages & Computing **19** (2008) 203–224
4. Yang, J., Fan, J., Hubball, D., Gao, Y., Luo, H., Ribarsky, W., Ward, M.: Semantic image browser: Bridging information visualization with automated intelligent image analysis. In: Visual Analytics Science And Technology, 2006 IEEE Symposium On, IEEE (2006) 191–198
5. Jing, Y., Rowley, H., Wang, J., Tsai, D., Rosenberg, C., Covell, M.: Google image swirl: a large-scale content-based image visualization system. In: Proceedings of the 21st international conference companion on World Wide Web, ACM (2012) 539–540
6. Barnard, K., Forsyth, D.: Learning the semantics of words and pictures. In: Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on. Volume 2., IEEE (2001) 408–415
7. Carson, C., Belongie, S., Greenspan, H., Malik, J.: Blobworld: Image segmentation using expectation-maximization and its application to image querying. Pattern Analysis and Machine Intelligence, IEEE Transactions on **24** (2002) 1026–1038
8. Schyns, P.G., Oliva, A.: From blobs to boundary edges: Evidence for time-and spatial-scale-dependent scene recognition. Psychological science **5** (1994) 195–200
9. Oliva, A., Torralba, A.: Modeling the shape of the scene: A holistic representation of the spatial envelope. International journal of computer vision **42** (2001) 145–175
10. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on. Volume 2., IEEE (2006) 2169–2178
11. Pass, G., Zabih, R.: Histogram refinement for content-based image retrieval. In: Applications of Computer Vision, 1996. WACV'96., Proceedings 3rd IEEE Workshop on, IEEE (1996) 96–102
12. Oliva, A.: Gist of the scene. Neurobiology of attention **696** (2005) 251–258
13. Greene, M.R., Oliva, A.: The briefest of glances the time course of natural scene understanding. Psychological Science **20** (2009) 464–472

14. Greene, M.R., Oliva, A.: Recognition of natural scenes from global properties: Seeing the forest without representing the trees. Cognitive psychology **58** (2009) 137–176
15. Castelhano, M.S., Henderson, J.M.: The influence of color on the perception of scene gist. Journal of Experimental Psychology: Human perception and performance **34** (2008) 660
16. Torralba, A., Oliva, A.: Statistics of natural image categories. Network: computation in neural systems **14** (2003) 391–412
17. Oliva, A., Torralba, A.: Building the gist of a scene: The role of global image features in recognition. Progress in brain research **155** (2006) 23–36
18. Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A.: Sun database: Large-scale scene recognition from abbey to zoo. In: Computer vision and pattern recognition (CVPR), 2010 IEEE conference on, IEEE (2010) 3485–3492
19. Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., Oliva, A.: Learning deep features for scene recognition using places database. In: Advances in neural information processing systems. (2014) 487–495
20. Chang, P., Krumm, J.: Object recognition with color cooccurrence histograms. In: Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on. Volume 2., IEEE (1999)
21. Huang, J., Kumar, S.R., Mitra, M., Zhu, W.J., Zabih, R.: Image indexing using color correlograms. In: Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on, IEEE (1997) 762–768
22. Van De Sande, K.E., Gevers, T., Snoek, C.G.: Evaluating color descriptors for object and scene recognition. Pattern Analysis and Machine Intelligence, IEEE Transactions on **32** (2010) 1582–1596
23. Bosch, A., Zisserman, A., Muoz, X.: Scene classification using a hybrid generative/discriminative approach. Pattern Analysis and Machine Intelligence, IEEE Transactions on **30** (2008) 712–727
24. Oliva, A., Schyns, P.G.: Diagnostic colors mediate scene recognition. Cognitive psychology **41** (2000) 176–210
25. Yang, J., Jiang, Y.G., Hauptmann, A.G., Ngo, C.W.: Evaluating bag-of-visual-words representations in scene classification. In: Proceedings of the international workshop on Workshop on multimedia information retrieval, ACM (2007) 197–206
26. Hou, J., Gao, H., Xia, Q., Qi, N.: Feature combination and the knn framework in object classification. IEEE transactions on neural networks and learning systems **27** (2016) 1368–1378
27. Swain, M.J., Ballard, D.H.: Color indexing. International journal of computer vision **7** (1991) 11–32
28. Tibshirani, R., Walther, G., Hastie, T.: Estimating the number of clusters in a data set via the gap statistic. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **63** (2001) 411–423