Learning global properties of scene images from hierarchical representations

Anonymous Author(s)
Affiliation
Address
email

Abstract

Scene images with similar spatial layout properties often display characteristic statistical regularities on a global scale. In order to develop an efficient code for these global properties that reflects their inherent regularities, we train a hierarchical probabilistic model to infer conditional correlational information from scene images. Fitting a model to a scene database yields a compact representation of global information that encodes salient visual structures with low dimensional latent variables. Using perceptual ratings and scene similarities based on spatial layouts of scene images, we demonstrate that the model representation is more consistent with perceptual similarities of scene images than the metrics based on the state-of-the-art visual features.

1 Introduction

Understanding the global structures in scene images (pictures that depict spaces rather than primarily describing objects in a scene) is a key process for holistic perception of scenes. Such global information gives rise to relevant perceptual spatial layout properties of scene images such as depth, opennes and perspective [6]. In addition, scene images that belong to the same categories tend to have similar global structures [15] suggesting that the global information contributes to semantic properties of scenes.

Previous studies have revealed that global features such as GIST [16], pyramid of histograms of orientation gradients (PHOG) [2], spatial pyramid of SIFT [11] and histograms of textons [4] are capable of predicting the semantic properties of scene images such as perceptual properties of the spatial layouts [19], categories, memorabiliy [8] and typicality [3] of scene images. Although these approaches have been successful, the features require careful tuning depending on the tasks.

Another potential disadvantage of projecting scene images onto hand-designed feature spaces is that they do not necessarily capture all relevant scene information. For instance, although scene images have diverse local properties based on their contents (textures and objects within the scenes, etc.), the global structures of scenes are highly constrained similarities in spatial layout and 3D structure. These constraints provide scene images with special regularities on the global scale. Hand-designed representations which do not take these regularities into account is unlikely to deal with the meaningful statistical structures of the scene images (which are potentially relevant to the perceptual properties of scene images) [17].

Several algorithms have been developed for encoding the characteristic structures of images. One approach is to build efficient representations that encode images with a small number of coefficients [21, 7]. Another method is to learn a representation invariant to translations and rotations [13, 10, 18]. This algorithm adopts pooling algorithms that feed the strongest responses of local filters over a fixed range to the higher level representations. Although these methods have been successful for

local textures and object recognition, scene images have quite different properties from them and thus such objectives might not be optimal.

For learning regularities of scene images, one interesting objective would be to encode the co-occurrences of local structures on global scales. For instance, horizontal lines, which are prevalently observed structures in scene images, are composed of horizontal structures over space around similar vertical locations. A model which can encode such prevalent global structures based on the co-occurrences of local structures would be able to represent global regularities of scene images. To learn a representation which is more adequate for the purpose of learning the global structures of the scene images, we train a hierarchical probabilistic model (which will be referred to hereafter as the distribution coding model) that infers the correlational structures of the distributions from which specific types of scenes are drawn [9]. The distribution coding model encodes scene images with latent variables that compactly represent the space of covariance matrices that best capture correlational structure of the scene mages. Since the model encodes a scene image based on its distribution but not its pixel values, it is invariant to image variability that is not aligned with the statistical regularities of scene images.

The contributions of this paper are : 1) we optimize the learning and inference procedures for the distribution coding model expediting the training process, 2) we put more sophisticated constraints on the model parameters than previous approach to prevent degenerate solutions, 3) the parameters of the distribution coding model fitted to scene images reveal global structures which are prevalent in scene images, 4) the latent variables for encoding the correlational structures of scene images compactly encode the perceptually salient visual structures of scene images, and 5) develop a scene similarity measure based on the distribution coding model which is significantly more consistent with perceptual similarities of scene images than state-of-the-art descriptors.

2 Model training

2.1 Model description

To learn the global structures captured by the correlational relationships over space, we trained DCM [9] on whole scene images. DCM assumes that a data point, $\mathbf{x}$, e.g., a vectorized scene image in our setting, follows a conditional multivariate gaussian distribution,

$$\mathbf{x}|\mathbf{y} \sim \mathcal{N}(0, \mathbf{C}(\mathbf{y})) \tag{1}$$

To satisfy the positive definiteness constraint on covariance matrices, the model formulates the logarithm of the covariance matrices as a function of the latent variable $\mathbf{y}$ as below,

$$\log(\mathbf{C}(\mathbf{y})) = \sum_j y_j \mathbf{A}_j = \sum_j y_j \sum_k w_{j,k} \mathbf{b}_k \mathbf{b}_k^T \tag{2}$$

where y_j corresponds to the j th element of $\mathbf{y}$. With this formulation, DCM is capable of defining a continuum of covariance matrices that are defined by the continuous latent variables $\mathbf{y}$. Note that the model encodes $\mathbf{x}$ in terms of its *distribution* unlike other scene descriptors. This approach makes the representation robust to noise which is not relevant to the regularities present in the scene images.

Since $\mathbf{A}_j$ is symmetric, DCM formulates it as the weighted sum of the outer products of vectors $\mathbf{b}_k$'s whose dimensionality is identical to that of the data. Each $\mathbf{b}_k$ corresponds to a direction along which the covariance matrices can vary in a manner analogous to eigenvectors. If we generate sample images using a multivariate Gaussian distribution with the covariance matrix $\exp(\mathbf{b}_k \mathbf{b}_k^T)$, the pixels located at the same positions as the elements of $\mathbf{b}_k$ which have the same signs will be correlated in the generated samples. On the other hand, if two elements of $\mathbf{b}_k$ have opposite signs, then the pixel values at the corresponding locations in the generated samples to these elements will be anti-correlated. Rather than learning separate sets of $\mathbf{b}_k's, (k = 1, \cdots, K)$ for each $\mathbf{A}_j$, the model lets them share the common dictionary of $\mathbf{b}_k$'s and incorporate coefficients $w_{j,k}$ to reduce the dimensionality of the parameters; $\mathbf{A}_j$ with a high value of $w_{j,k}$ strongly encodes the correlational structures present in $\mathbf{b}_k$. On the other hand, a low value of $w_{j,k}$ corresponds to a suppressed variability along $\mathbf{b}_k$. We constrain $\mathbf{b}_k$ and $\mathbf{w}_j = [w_{j,1}, \cdots, w_{j,K}]$ on the unit norm ball to prevent degenerate solutions [1].

To enforce the model parameters to learn a sparse and independent representation of covariance matrices, the model uses a Laplacian prior on $\mathbf{y}$,

$$\log p(\mathbf{y}) \propto -\sum_j |y_j| \tag{3}$$

2.2 Learning and inference

The model parameters $\Theta = \{\mathbf{b}_k, \mathbf{w}_j\}$ were optimized using the maximum likelihood method. During the training process, we randomly sample a subset of data. We first infer latent variables for each data point in the subsample with $\mathbf{b}_k$'s and $\mathbf{w}_j$'s fixed to the current estimation (inference step). Then, with the latent variables fixed, we update the model parameters (learning step).

Once the training process is completed, we can use the model parameters $\mathbf{b}_k$'s and $\mathbf{w}_j$'s to infer the latent variables for new scene images. We do so by using the same procedure that we used in the inference step in the training process. Latent variables are initialized to random and updated to maximize the likelihood of a scene image.

The number of $\mathbf{b}_k$'s and the number of $\mathbf{w}_j$'s, K and J, are fixed beforehand (here, K = 596 and J = 60). Note that manipulating the two parameters for DCM is not exactly comparable to tuning the parameters for hand-engineered features such as GIST and HOG, but rather more analogous to choosing the number of principal components in PCA; as you increase the number of parameters, the model considers the noisier part of the distribution and thus the model is not so sensitive to the parameter values after you reach a certain number.

2.3 Optimization method

The previous implementation of DCM [9] employed the stochastic gradient method for the learning and inference procedures. While the stochastic gradient method is easy to implement, the method requires sophisticated tuning of the learning parameters such as step sizes. Here, we adopt the limited memory BFGS (L-BFGS) method [14] for the learning and inference procedures. Since the L-BFGS method employs the line search method to find the step sizes, there is no need to tune them. Another benefit of the L-BFGS method is that it approximates the second-order information and thus converges faster with greater stability than the stochastic gradient method. To handle the large size of the dataset required for estimating the high dimensional parameters, we trained the model with the minibatch training method [12].

With the optimized learning procedures, the model converges within hours to a good solution whereas the previous implementation took days. The results we report in this paper were obtained with approximately 20 hours of the learning procedure on a GPGPU Tesla M2070 GPU. Note that once we fit the model parameters through the training procedure, extracting features from images which corresponds to the inference step is achieved very quickly.

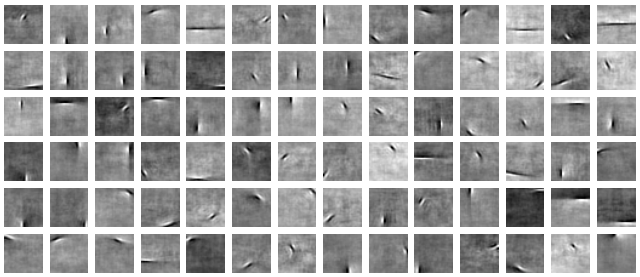
2.4 Training data and preprocessing

We trained DCM on 130,519 scene images (from 397 scene categories) in the SUN database [22]. The dataset is hierarchically organized and covers wide varieties of scene images with diverse structures. Due to the technical constraints such as the number of training examples required for avoiding overfitting and the computational cost, we downsampled the original scene images to 32×32 grayscale images suitable for object detection and scene categorization tasks by human subjects [20]. Because the dataset has sufficient number of scene images compared to the dimensionality of the model parameters, it is unlikely that the results are overfitted to the training data. This is demonstrated when we apply the model parameters trained on the SUN database to other scene image datasets [11, 19] and scene images downloaded from the web, as the latent variables have similar properties.

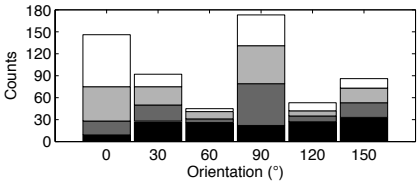
3 Model representation

3.1 Model parameters

As discussed in Section 2.1, $\mathbf{b}_k$ encodes a common direction along which the covariance units $\mathbf{A}_j$'s can vary. When trained on the 32×32 scene images, $\mathbf{b}_k$'s show gabor-like structures as shown in Figure 1a. Note that the formulation of the model did not constrain $\mathbf{b}_k$'s to have localized structures; rather, the structures automatically emerged while fitting the parameters to the scene image statistics.



(a) A subset of $\mathbf{b}_k$'s



(b) Orientations and scale distribution of $\mathbf{b}_k$'s

Figure 1: (a) 84 out of 576 are shown. To visualize $\mathbf{b}_k$'s which are vectors, we rearrange their elements into 32×32 matrix form. More examples are shown in Supplementary Figure 2. (b) The stacked histogram describing the orientation and scale of $\mathbf{b}_k$'s. 0° corresponds to the horizontal orientation, 90° to the vertical orientation. The $\mathbf{b}_k$'s are sorted from the most localized to the most global based on their scales. The black, dark gray, light gray and white parts of the bar graph correspond respectively to the group of the top 25% localized structures, the groups of top 25–50% and 50–75% localized $\mathbf{b}_k$'s and the group of the most global $\mathbf{b}_k$'s.

When we categorize $\mathbf{b}_k$'s based on their orientation and scale, the horizontal and vertical orientations are dominant in light of the external physical structures [5]. In terms of scale, horizontal units, compared to other orientations, have a greater portion of the most global scales (Figure 1b). The non-isotropic distribution of scale and orientation of $\mathbf{b}_k$'s suggests DCM invests more resources for prevalent visual structures in scene images. This contrasts with most hand-designed visual features in that they tend to allocate uniform bits of information for all orientations and scales.

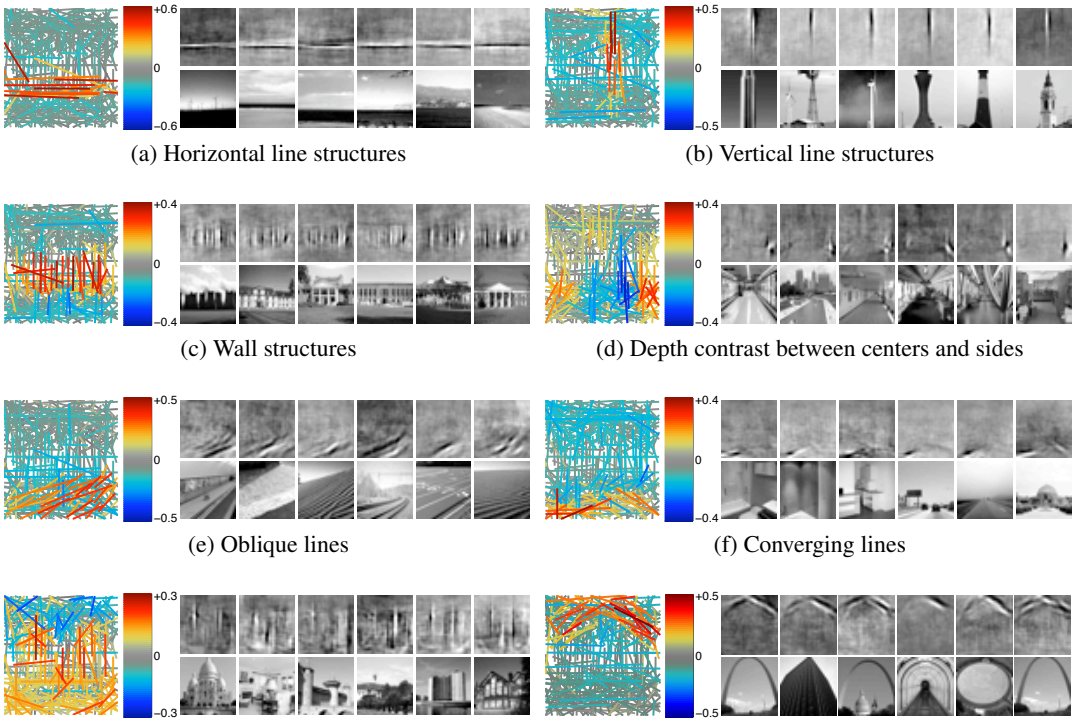
While $\mathbf{b}_k$'s have localized properties, we find that $\mathbf{w}_j$'s encode global information by incorporating the localized correlational structures encoded in the $\mathbf{b}_k$'s over space. To visualize each $\mathbf{w}_j$, we first assign a bar to each $\mathbf{b}_k$ which has the same location, orientation and scale with that $\mathbf{b}_k$ in the image space. We then assign each bar a color value corresponding to the value of $w_{j,k}$ [9]. We show six out of sixty $\mathbf{w}_j$'s, equivalently $\mathbf{A}_j$ (Eq.2), in Figure 2; these $\mathbf{w}_j$'s reveal horizontal and vertical line structures (Fig.2a–2b), wall structures (Fig.2c), contrasts between centers and sides (Fig.2d), converging lines (Fig.2e) and contrasts between upper and lower parts (Fig.2f).

We demonstrate the global correlational structures encoded in $\mathbf{w}_j$ by generating random samples from a multivariate Gaussian distribution whose covariance matrices is $\exp(y_j \mathbf{A}_j)$ ($y_j > 0$). The generated samples show visually similar structures as the corresponding covariance matrices. In addition, scene images which have the highest values of y_j among the SUN database contain visual structures that resemble the visualization of correlational structures encoded in $\mathbf{A}_j$.

3.2 Latent variables

Due to the sparsity constraint on the latent variables (Eq.3), the distribution of latent variables $\hat{y}$ peaks around zero. Even though there are 60 covariance units ($\mathbf{A}_j$), only approximately 20 units are necessary for capturing the correlational structures of a scene image; when we order the elements of the latent variable $\hat{y}$ of a scene image $\mathbf{x}$ according to their magnitudes, and maintain the values of the most active elements, while setting others to zero to compute the likelihood of $\mathbf{x}$, the log likelihood is saturated when we use 20 most active units. Note that this number corresponds to only less than 2% of the original dimensionality of 32×32 grayscale images.

When we visualize the covariance matrices determined by the latent variables, they are visually similar to the salient visual features of the corresponding scene images (Figure 3). For each sample scene image, we order its latent variables $\hat{y} = \{\hat{y}_1, \cdots, \hat{y}_J\}$ based on their magnitudes. We show the logarithms of the cumulative covariance matrices, $\sum_{i=1}^{k} \hat{y}_{I(i)} \mathbf{A}_{I(i)}$, in the first rows; I corresponds to the order of $\hat{y}_j$s based on the absolute values in the descending order. The positive and negative components of $\hat{y}_{I(k)} \mathbf{A}_{I(k)}$ are separately displayed in the second and the third rows separately for visual clarity. The second column corresponds to $k = 1$ and the right-most column corresponds to $k = 6$. Consistent with the sparse distribution of $\hat{y}$, the first few elements of the $\hat{y}_j$ encode the salient global structures of scene images.



(a) Horizontal line structures (b) Vertical line structures

(c) Wall structures (d) Depth contrast between centers and sides

(e) Oblique lines (f) Converging lines

(g) Contrast between top and bottom (h) Structures in upper parts

Figure 2: (a)–(h) Representative $\mathbf{A}_j$ on the left with corresponding color bars. The red corresponds to positive values of $w_{j,k}$ while blue represents the negative values. On the right, top rows show images generated from multivariate Gaussian distributions with $\exp(y_j \mathbf{A}_j)$ as covariance matrices ($y_j > 0$). The bottom rows show scene images from the SUN database which have the highest values of $\hat{y}_j$.

4 Quantitative evaluation of DCM on spatial layout representation

4.1 Similarity measure based on the distribution coding model

Once we train DCM and infer the latent variables for scene images, we can develop a metric for measuring the scene similarities in terms of correlational structures between two scene $\mathbf{x}_t$ and $\mathbf{x}_c$ as below:

$$d(\mathbf{x}_t, \mathbf{x}_c) = -\log p(\hat{y}_c|\mathbf{x}_t) - \log p(\hat{y}_t|\mathbf{x}_c) = -\log(p(\mathbf{x}_t|\hat{y}_c)p(\hat{y}_c) - p(\mathbf{x}_t)) - \log(p(\mathbf{x}_c|\hat{y}_t)p(\hat{y}_t) - p(\mathbf{x}_c)) \quad (4)$$

where $p(\mathbf{x}_c)$ is approximated by $p(\mathbf{x}_c|\hat{y}_c)p(\hat{y}_c)$ due to the intractability of evaluating the full integral [9]. The metric $d(\mathbf{x}_t, \mathbf{x}_c)$ is greater than zero and is equal to zero if $\mathbf{x}_t$ and $\mathbf{x}_c$ are identical. If two data points, $\mathbf{x}_t$ and $\mathbf{x}_c$, have similar correlational structures, then $\mathbf{x}_t$ will be highly likely under the multivariate Gaussian distribution with the covariance matrix determined by the latent variable for $\mathbf{x}_c$ and vice versa; thus the conditional probability of $\mathbf{x}_t$ given $\hat{y}_c$, $p(\mathbf{x}_t|\hat{y}_c)$ (Eq. 1), and also $p(\mathbf{x}_c|\hat{y}_t)$ will be high resulting in the low value of $d(\mathbf{x}_t, \mathbf{x}_c)$. On the other hand, if the two data points do not share similar correlational structures, $d(\mathbf{x}_t, \mathbf{x}_c)$ will have a high value. Due to the zero mean assumption in the model, we normalize the joint probability by the marginal probability of the images under the model assumption.

We demonstrate the usage of the similarity measure based on DCM above using the image retrieval task; for a target image $\mathbf{x}_t$, we retrieve candidate scene images from the SUN database. In Fig. 4, we show the five most similar candidate scene images with DCM GIST, HOG, PHOG and spatial pyramid of SIFT. For DCM we used the similarity measure defined in Eq. 4 and for others we used the Euclidean distances between features. For GIST and HOG, we tried three different spatial scales (1×1, 2×2 and 4×4) and show the qualitatively best results. Even though the model representa-

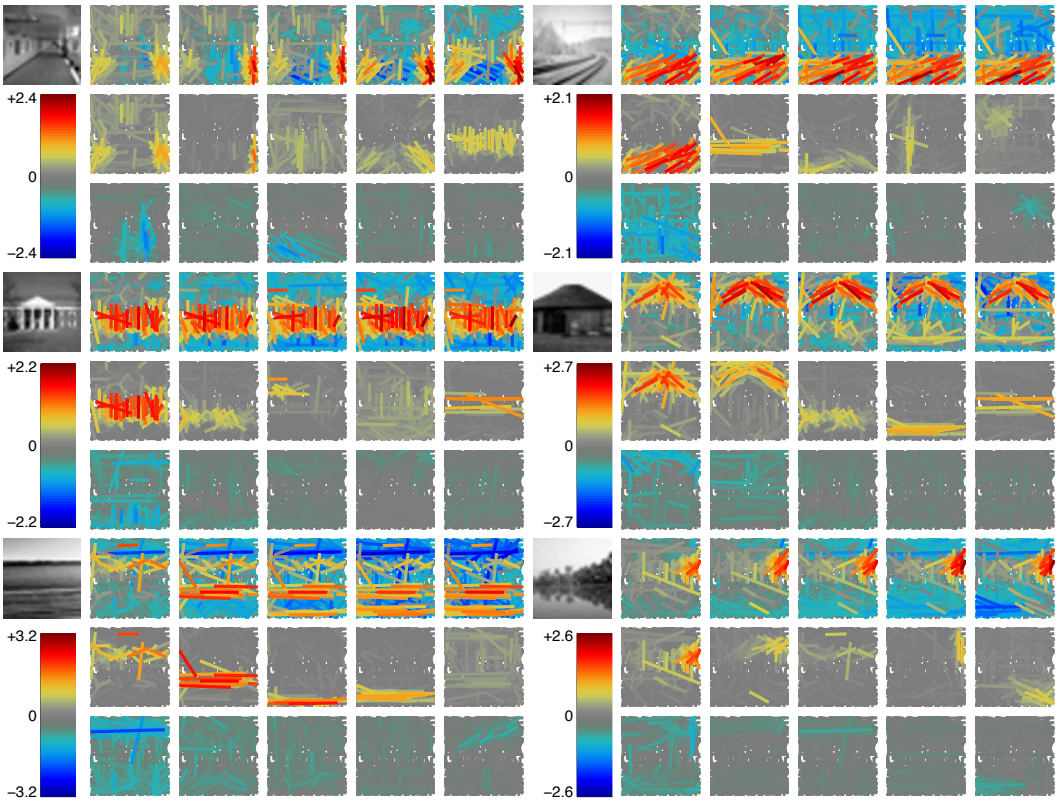


Figure 3: For each target image $\mathbf{x}$, we infer its latent variable $\hat{y}$ (Section 2.2) and order the $\hat{y}_j$s according to their absolute values. The first rows show the cumulative sum of the logarithm of the covariance matrix using the k most active $\hat{y}_j$s. The second and the third rows show the positive and negative parts of $\hat{y}_{I(k)}\mathbf{A}_{I(k)}$, respectively. I refers to the order of $\hat{y}_j$s based on their magnitudes. This figure is best viewed in color.

tion requires just a small number of units to represent a scene image, the similarity results show satisfactory results.

4.2 Perceptual similarities of scene images based on spatial layouts

To investigate whether the global correlational information encoded by DCM is consistent with the perceptual similarities between scene images, we conducted a perceptual experiment in which subjects were asked to select candidate scene images that were most similar to a target image in terms of spatial layouts. For each target image in the SUN database, the candidate images in the SUN database were chosen by retrieving the most similar image in the same semantic category to the target image based on various feature representations (DCM, GIST, HOG, PHOG and Spatial pyramid representations) and two levels of resolutions (32×32 and 128×128 except for DCM). All stimuli were displayed in 128×128 resolution and we down sampled the stimuli to 32×32 resolution for extracting features for 32×32 condition. For DCM, we used the probabilistic distance measure in Eq. 4 as the similarity measure and for other representations the similarities were computed by the Euclidean distances between the features.

Subjects were allowed to select more than one candidate images if they were equally similar to the target images. In the trials when none of the candidate images were perceptually similar to the target images or when the target images mainly consisted of objects and it was thus difficult to get a sense of spatial layout from the them, subjects could skip the trial. Subjects were specifically instructed to focus on the shape and spatial layout of the scenes and to ignore non-spatial attributes such as color or types of objects in the scenes. Using candidate images from the same category prevents subjects from depending on any semantic information to perform the task. Six subjects (one female; with normal or corrected to normal vision) participated in the experiment. We collected 3870 trials and the subjects selected 1.31 candidate images per trial on average (the number of candidate images

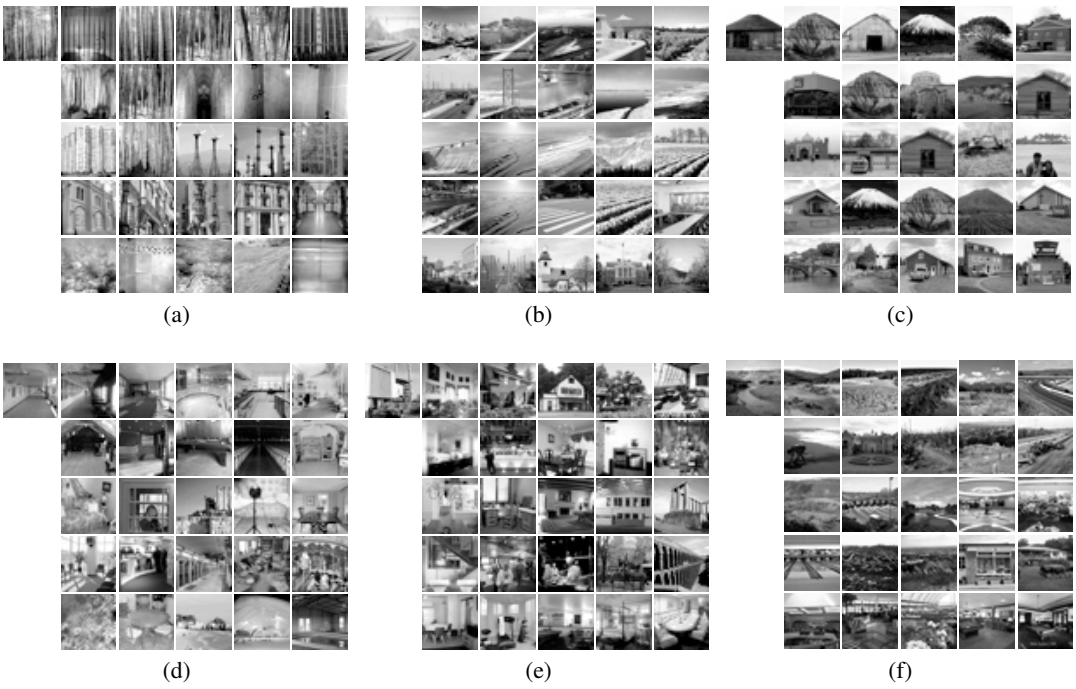


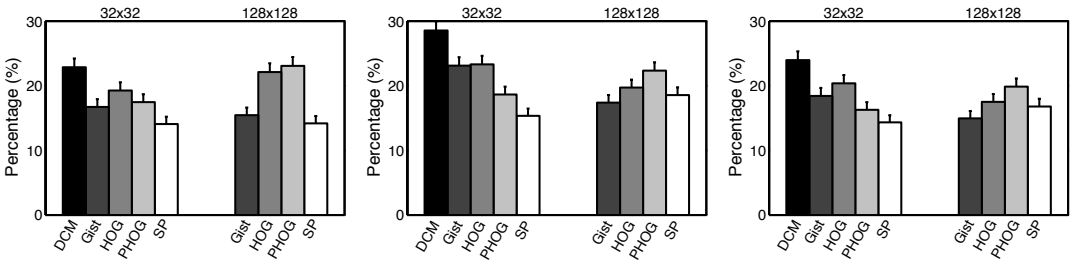
Figure 4: (a)–(f) Scene image retrieval results. The top left portion shows the target scene images. The retrieved images are ordered so that the left-most columns show the most similar and the right-most columns show the 5 th similar candidate scene images to the targets. From the top to the bottom rows correspond to DCM, GIST(4×4), HOG(2×2), PHOG (3 levels), spatial pyramid of SIFT (3 levels).

selected per trial ranged from 0 to 9). Out of 3870 trials, subjects selected more than one similar images in 1242 trials and selected zero similar images in 1213 trials. Trials in which subjects selected zero candidate image were discarded for further analysis.

We evaluate various similarity measures based on the percentage of trials in which candidate images retrieved by different representations were selected by the observers to be most similar to the corresponding target images. Interestingly, the performance of various representations show different patterns for the three different top-level categories in the SUN database: indoor, outdoor manmade (outdoor scenes with artificial structures such as buildings) and outdoor natural scenes. DCM outperforms other representations for outdoor natural and outdoor manmade scenes (Student's t-test; $p < 0.05$), but show comparable performance to PHOG and HOG for indoor scenes. We speculate that DCM seem not to be able to encode sharp edge information (as illustrated in the spectral analysis of Figure 3), making the representation less optimal for indoor scenes. However, for outdoor scenes whose spatial layout properties are less dependent on edge information, the compact representation of global structures of scene images by DCM are more consistency with perception.

4.3 Perceptual spatial layout ratings

As we demonstrated in Figure 4, the correlational patterns between the linear filter outputs show varying patterns for scene images with different perceptual spatial layout properties. In this section, we quantitatively evaluate how well the correlational information encoded by DCM predicts the perceptual ratings of mean depth and openness collected on a continuous 1-to-6 scale [19]; we do so by evaluating if two scene images which are similar based on a metric have similar perceptual ratings. Openness of a scene refers to the quantity and location of boundaries in a scene (1= large portion of unobstructed sky and dominant horizontal lines; 6= closed scenes, which have limited spatial extent). Mean depth refers to depth in a global sense related to the physical size of a scene (1= close to the camera, 6= far).



(a) Indoor (b) Outdoor Natural (c) Outdoor Manmade

Figure 5: Percentage of trials in which scene images retrieved by various representations were selected by the subjects. Error bars represent standard errors. DCM (distribution coding model), GIST (2×2), HOG (2×2), PHOG (L=2) and Spatial Pyramid (L=2). (a) Indoor, (b) Outdoor natural, (c) Outdoor manmade.

Table 1: AUCs for the image retrieval task based on scene layout ratings from [19]

Feature	Resolution	Depth			Openness		
		r=0.05 ($\times 10^{-2}$)	r=0.1 ($\times 10^{-2}$)	r=1 ($\times 10^{-1}$)	r=0.05 ($\times 10^{-1}$)	r=0.1 ($\times 10^{-1}$)	r=1 ($\times 10^{-1}$)
DCM	32×32	**4.16**	**7.53**	**5.90**	**1.11**	**1.41**	**5.24**
GIST (8×8)	256×256	4.13	7.49	5.88	1.05	1.34	4.00
ICA	32×32	3.89	7.10	5.61	0.64	0.86	3.95
HOG (4×4)	256×256	4.00	7.20	5.74	1.03	1.32	3.98
PHOG (L=3)	256×256	4.06	7.35	5.80	1.30	1.30	3.98
SIFT (L=3)	256×256	4.10	7.40	5.82	0.81	1.06	3.83

For each image, we set images with similar perceptual rating values as positive examples and others as negative. We use three different levels for the definition of positive examples; r=0.05, 0.1 and 1. In r=0.05 condition, for each image with assigned perceptual rating values, we set images which differ less than 0.05 in its perceptual rating to the image as its positive examples. We evaluate each representation with area under the curve (AUC) of the precision and recall curves. As in the previous section, we used the similarity measure defined in Eq. 4 for DCM and the Euclidean distances between features for other representations. DCM has higher AUC values for both mean depth and openness at all three ranges tested (Student's t-test; $p < 0.01$) (Table 1). Consistent with the results from the previous section, this result suggests that the global structures that DCM automatically learns from the scene images effectively encode perceptually relevant information.

5 Conclusion

We trained DCM to learn the correlational information on the whole scene images. The model parameters show global correlational structures reflecting the regularities inherent in the scene images. Adaptive representation to the characteristic statistics allows encoding of the data with a small number of latent variables. In addition, the experiment for perceptual scene image similarities suggest that the model representation is a good scene image descriptor with greater consistency with perceptual properties of the global structures. The probabilistic correlational distance developed in this paper can be used for image retrieval systems. Our approach can be extended to larger size images for encoding more detailed information by first learning the correlational structures on local patches and integrating the local information over space. Also, the probabilistic distance measure introduced in this paper can be utilized not only for whole image retrieval but also for finding local interest matching points between images.

References

[1] P. Absil, R. Mahony, and R. Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2008.

[2] A. Bosch, A. Zisserman, and X. Munoz. Representing shape with a spatial pyramid kernel. In *ACM International Conference on Image and Video Retrieval*, 2007.

[3] K. Ehinger, J. Xiao, A. Torralba, and A. Oliva. Estimating scene typicality from human ratings and image features. In *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*, 2011.

[4] L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *booktitle="IEEE Conference on Computer Vision and Pattern Recognition (CVPR)"*, pages 524–531, 2005.

[5] A. Girshick, M. Landy, and E. Simoncelli. Cardinal rules: visual orientation perception reflects knowledge of environmental statistics. *Nature Neuroscience*, 14(7):926–932, 2011.

[6] M. Greene and A. Oliva. Recognition of natural scenes from global properties: Seeing the forest without representing the trees. *Cognitive Psychology*, 58(2):137 – 176, 2009.

[7] A. Hyvärinen and P. Hoyer. Emergence of phase- and shift-invariant features by decomposition of natural images into independent feature subspaces. *Neural Computation*, 12(7):1705–1720, Jul 2000. doi: 10.1162/089976600300015312.

[8] P. Isola, J. Xiao, A. Torralba, and A. Oliva. What makes an image memorable? In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 145–152, 2011.

[9] Y. Karklin and M. Lewicki. Emergence of complex cell properties by learning to generalize in natural scenes. *Nature*, 457:83–86, January 2009.

[10] K. Kavukcuoglu, M. Ranzato, R. Fergus, and Y. LeCun. Learning invariant features through topographic filter maps. In *Proc. International Conference on Computer Vision and Pattern Recognition (CVPR'09)*. IEEE, 2009.

[11] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2:2169 – 2178, 2006.

[12] Q. Le, J. Ngiam, A. Coates, A. Lahiri, B. Prochnow, and A. Ng. On optimization methods for deep learning. In *In Proceedings of the Twenty-Eighth International Conference on Machine Learning*, 2011.

[13] H. Lee, Y. Largman, P. Pham, and A. Y. Ng. Unsupervised feature learning for audio classification using convolutional deep belief networks. In *Advances in Neural Information Processing Systems 22*, pages 1096–1104. 2009.

[14] D. Liu and J. Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical Programming*, 45(1):503–528, 1989.

[15] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42:145–175, Jan 2001.

[16] A. Oliva and A. Torralba. Building the gist of a scene: the role of global image features in recognition. volume 155, Part B of *Progress in Brain Research*, pages 23 – 36. Elsevier, 2006.

[17] B. Olshausen and D. Field. Natural image statistics and efficient coding. In *Network: Computation in Neural Systems, 7:333–339*, pages 333–339, 1996.

[18] M. Ranzato, F.-J. Huang, Y.-L. Boureau, and Y. LeCun. Unsupervised learning of invariant feature hierarchies with applications to object recognition. In *Proc. Computer Vision and Pattern Recognition Conference (CVPR'07)*. IEEE Press, 2007.

[19] M. Ross and A. Oliva. Estimating perception of scene layout properties from global image features. *Journal of Vision*, 10:1–25, 2010.

[20] A. Torralba. How many pixels make an image? *Visual neuroscience*, 26:123–131, Jan 2009.

[21] J. H. van Hateren and D. L. Ruderman. Independent component analysis of natural image sequences yields spatio-temporal filters similar to simple cells in primary visual cortex. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 265(1412):2315–2320, 1998.

[22] J. Xiao, J. Hays, K. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 0:3485–3492, 2010.