Learning global properties of scene images based on their correlational structures

Wooyoung Lee
Machine Learning Department
Carnegie Mellon University
Pittsburgh, PA 15213
wooyoung@cs.cmu.edu

Michael S. Lewicki
EECS
Case Western Reserve Univesity
Cleveland, OH
mike.lewicki@case.edu

Abstract

Scene images share underlying regularities on the global scale. In order to develop a representation that encodes the global properties of scene images and reflects their inherent regularities, we train a probabilistic hierarchical model to infer correlational information from scene images. The model parameters fitted to the characteristic statistics of scene images reveal an efficient representation of global information that encodes salient visual structures with low dimensional latent variables. Through a perceptual experiment which assesses scene image similarities in terms of spatial layout, we demonstrate that our model representation is more consistent with perceptual similarities of scene images than 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 [5]. In addition, scene images that belong to the same semantic categories tend to have similar global structures [14] suggesting that the global information contributes to semantic properties of scenes.

Previous studies have revealed that global features such as GIST [15], pyramid of histograms of orientation gradients (PHOG) [2], spatial pyramid of SIFT [10] 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, memorability [7] and typicality [3] of scene images. Although these approaches have been successful, the features require careful hand-tuning of parameters depending on the tasks. This requirement limits the generality of what is learned based on such features in one dataset to others [21].

Another potential disadvantage of projecting scene images onto the 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 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) [16]. Therefore, such representations require extra procedures such as supervised training or metric learning with rather expensive human labels in order to learn proper metrics relevant to higher level representation such as perception or semantic categories [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 by imposing sparsity constraints [22, 6]. Another method is to learn a representation invariant to translations and rotations [12, 9, 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 [8]. The distribution coding model compactly represents 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 that : 1) we show a compact dictionary for representing global structures of scene images, 2) the latent variables for encoding the correlational structures of scene images compactly encode the perceptually salient visual structures of scene images, 3) we 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, 4) we optimize the learning and inference procedures for the distribution coding model expediting the training process and 5) we put more sophisticated constraints on the model parameters than previous approach to prevent degenerate solutions.

2 Model training

2.1 Model description

To learn the global structures captured by the correlational relationships over space, we trained the distribution coding model [8] on whole scene images. The distribution coding model assumes that data, $\mathbf{x}$, e.g., vectorized scene images in our setting, follows a conditional multivariate gaussian distribution,

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

The zero mean assumption is valid because averaging a sufficient number of scene images shows that the pixel values of the mean scene image have almost uniform values. 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, the distribution coding model 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, the distribution coding model 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. Rather than learning separate sets of $\mathbf{b}_k, (k = 1, \cdots, K)$ 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 compact 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 training data. We first infer latent variables for each data points 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 which are fit to the statistical properties of scene images 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. Based on the formulation of the distribution coding model, the latent variables optimized for a scene image compactly encode the covariance matrix for the multivariate Gaussian distribution from which the scene image is drawn.

The number of $\mathbf{b}_k$s and the number of $\mathbf{w}_j$s, K and J, are fixed beforehand. The results that we report in this paper were obtained with the K and J set to 596 and 60, respectively.

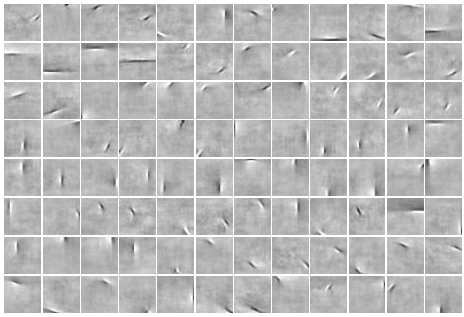
2.3 Optimization method

The previous implementation of the distribution coding model [8] 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 [13] 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 deal with the large size of the dataset required for estimating the high dimensional parameters, we trained the model with the minibatch training method [11].

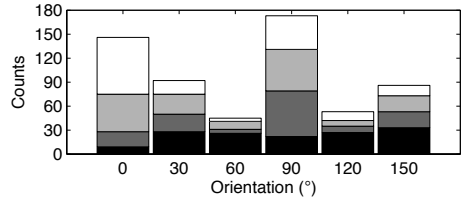
With the optimized learning procedures, the model converges within hours to a good solution whereas the previous implementation took days to reach stable solutions. The results we report in this paper were obtained with approximately 20 hours of the learning procedure on a GPGPU Tesla M2070 GPU. Note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

2.4 Training data and preprocessing

We trained the distribution coding model on 130,519 scene images (from 397 scene categories) in the SUN database [23]. 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 performance of object detection and scene categorization tasks by human subjects [20]. Because the dataset has enough 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 [10, 19] and scene images downloaded from the web, as the latent variables have similar properties.



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



(b) Stacked histogram of orientations and scales of $\mathbf{b}_k$s

Figure 1: (a) 96 out of 576 randomly selected are shown. To visualize $\mathbf{b}_k$s which are vectors, we rearrange their elements into 32×32 matrix form. (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.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.

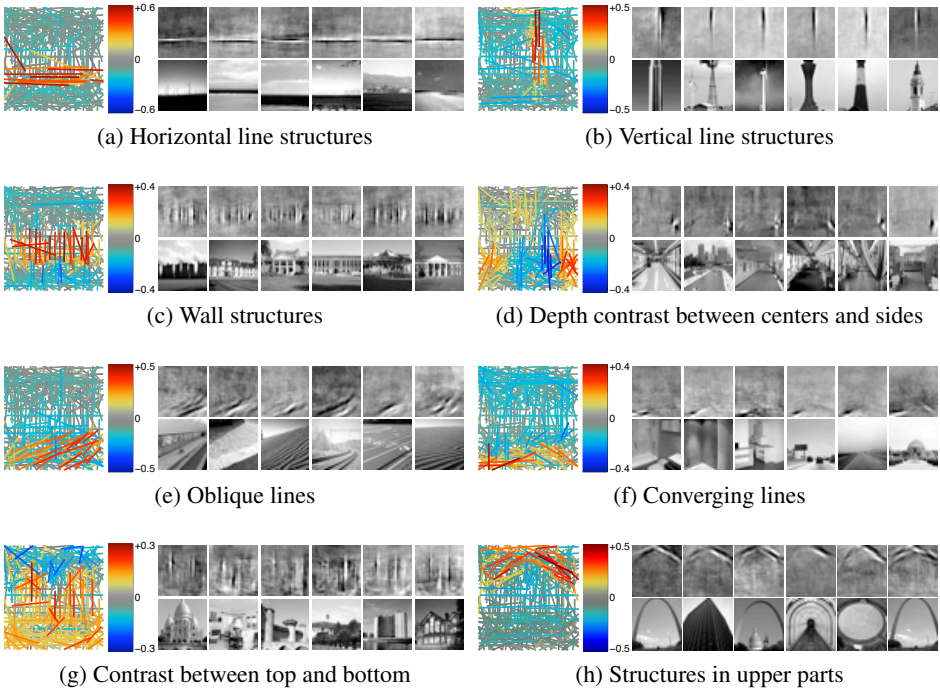
3 Model representation

3.1 Model parameters

As discussed in Section 2.1, $\mathbf{b}_k$ encodes a common direction along which the covariance units $\mathbf{A}_j$ 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 emerged while fitting the parameters to the scene image statistics. 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 found at the same location with theses elements in the generated samples will be anti-correlated.

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. 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, the common directions along which the covariance units $\mathbf{A}_j$s can vary, suggests the density component model 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 showed 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 and orientation with that $\mathbf{b}_k$ in the image space. We then assign each bar a color value cor-



(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

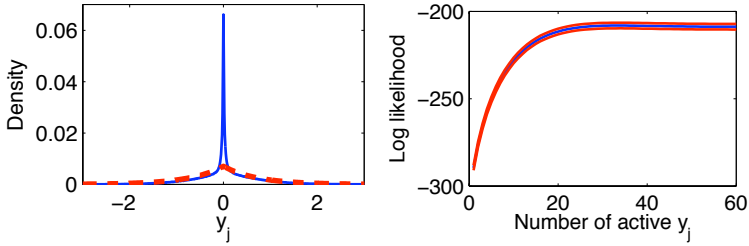
(h) Structures in upper parts

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 $\hat{y}_j$.

responding to the value of $w_{j,k}$. We show eight out of sixty $\mathbf{w}_j$s, equivalent to the $\mathbf{A}_j$ (Eq.2) in Figure 2; these $\mathbf{w}_j$ reveal horizontal and vertical line structures (Fig. 2a–2b), wall structures (Fig. 2c), depth contrasts between centers and sides (Fig. 2c), oblique lines (Fig. 2e), converging lines (Fig. 2f), contrasts between top and bottom (Fig. 2g) and structures in upper part of images (Fig. 2h). We demonstrate the global correlational structures encoded by $\mathbf{w}_j$ 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 (Fig. 3a). Even though there exist 60 covariance units ($\mathbf{A}_j$), only approximately 20 units are necessary for capturing the correlational struc-



(a) Distribution of $\hat{y}_j$ (b) Log likelihood

Figure 3: (a) Distribution of values of $\hat{y}_j$ for scene images in the SUN database (blue solid line) and the constraint we imposed on y_j (Eq.3, red dashed line). (b) The log likelihood computed using the most active $\hat{y}_j$s. The x-axis corresponds to the number of most active units used (60 indicates using the original $\hat{y}$), while the y-axis corresponds to the log likelihood of the data computed using the most active $\hat{y}_j$s. The blue line corresponds to the mean over the SUN database and the red lines are the error bars.

tures of a scene image (Fig. 3b); 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 4). 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.

We can also analyze the covariance matrices that best describes corresponding scene images by spectral analysis. The spectral analysis reveals the directions along which the covariance matrices are expanded or contracted. In Figure 5, we visualize the eigenvectors of the covariance matrices. Note that the eigenvectors corresponding to the positive values of eigenvalues have similar global structures to the scene images. Also, the structures encoded in the eigenvectors corresponding to the negative values of the eigenvalues are absent in the corresponding scene images. Consistent with the previous analysis, this results suggest that the directions along which the covariance matrices are extracted encode the global structures of scene images.

Lastly, we show randomly generated samples drawn from multivariate Gaussian distributions with covariance matrices parameterized by latent variables corresponding to target scenes, respectively (Figure 6). It is interesting to note that the generated samples only preserve global structures corresponding to low frequency information. Note that

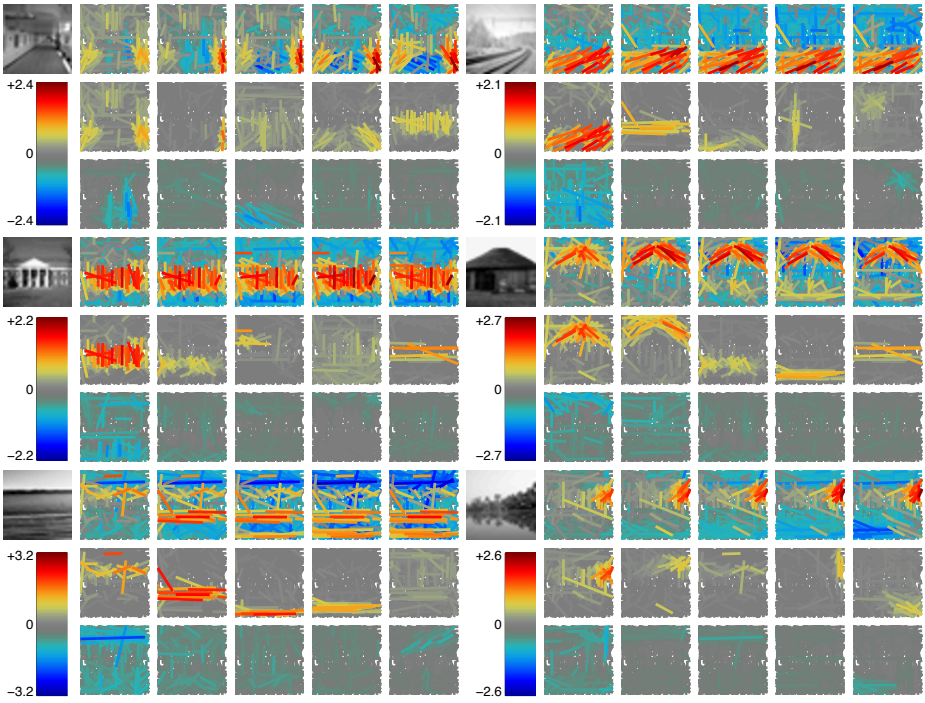


Figure 4: For each target image $\mathbf{x}$, we infer its latent variable $\hat{y}$ (Section 2.2) and order the $\hat{y}_j$s according to their absolute values. The first rows show the cumulative sum of the logarithm of the covariance matrix using the k most active $\hat{y}_j$s. The second and the third 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.

the generated samples, however, do not preserve the edges present in the original images and suggests that the covariance structures in images do not necessarily preserve contours.

4 Similarity measure based on the distribution coding model

In the previous section, we showed that the latent variables $\hat{y}$ capture the correlational information which is consistent with the visual structures of scene images and that this representation is efficient in that it requires only a small number of variables to encode the salient properties of scene images. In this section, we discuss how we can utilize the correlational structures encoded in the latent variables as a scene similarity measure and show image retrieval results based on it.

Once we train the distribution coding model and infer the latent variables for scene images, we can develop a metric for measuring the scene similarities in terms of correlational structures using the joint probability of a target scene image $\mathbf{x}_t$ and a latent

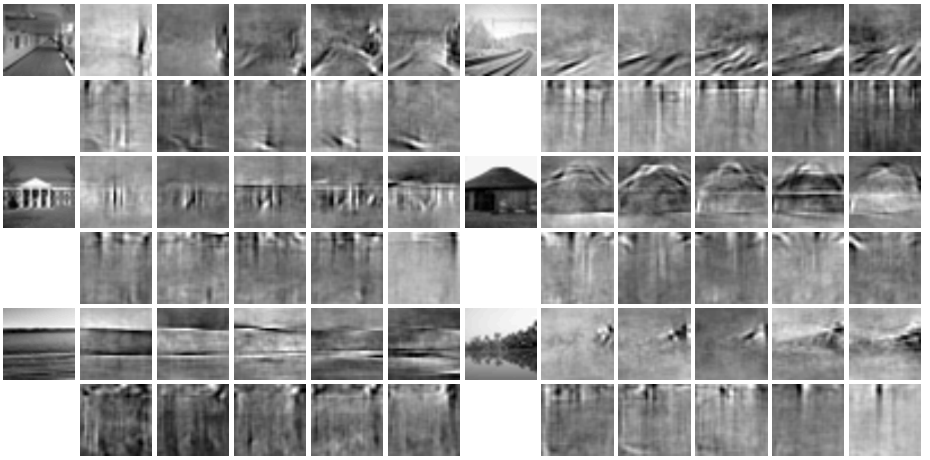


Figure 5: For each target image $\mathbf{x}$, we show eigenvectors corresponding to the positive values (upper row) and the negative values (lower row) of the eigenvalues.

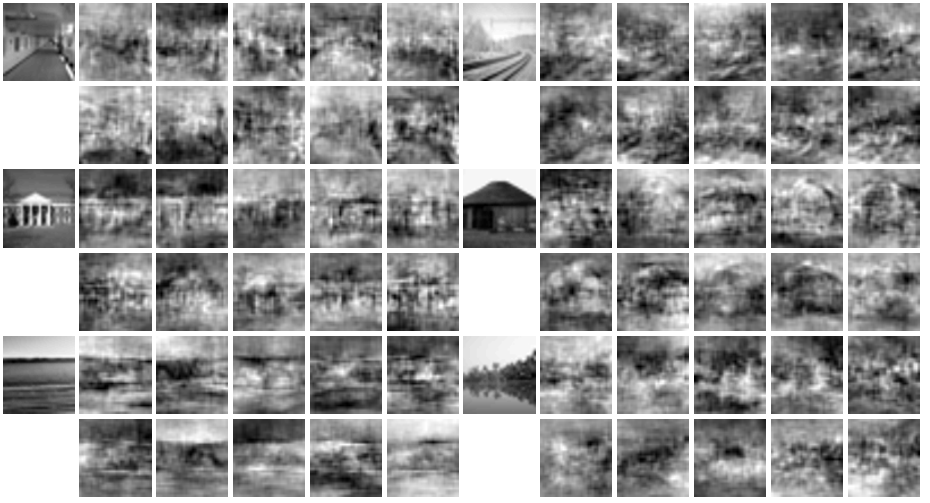


Figure 6: Generated random samples from multivariate Gaussian distributions with the covariance matrices parameterized by latent variables corresponding to original images.

variable $\hat{y}_c$ of a candidate scene image $\mathbf{x}_c$,

$$p(\mathbf{x}_t, \hat{y}_c) \propto p(\mathbf{x}_t|\hat{y}_c)p(\hat{y}_c) \tag{4}$$

The metric consists of two terms; the first term indicates the level of similarity between a target scene image and a candidate scene image in terms of correlational structures. 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 $\mathbf{x}_c$; thus the conditional probability of $\mathbf{x}_t$ given $\hat{y}_c$,

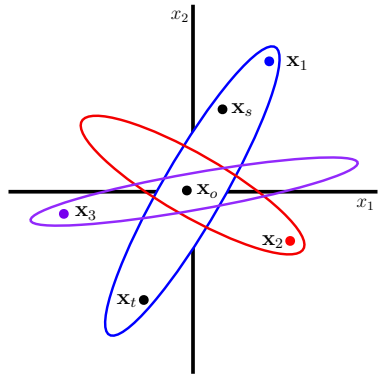


Figure 7: Schematic representation for $p(\mathbf{x}_t|\hat{y})$. The blue oval represents the covariance matrix $\mathbf{C}(\hat{y}_1)$ (Eq.3) where $\hat{y}_1$ indicates the latent variable for $\mathbf{x}_1$. The red oval represents the covariance matrix that captures anti-correlated x_1 and x_2 values. Under the Gaussian distribution with this covariance matrix, $\mathbf{x}_t$ will have low likelihood. The purple oval optimized for $\mathbf{x}_3$ represents the positively correlated values of x_1 and x_2, but to a different degree from x_1^t and x_2^t. Thus, $\mathbf{x}_t$ will have low conditional probability under the distribution optimized for $\mathbf{x}_3$.

$p(\mathbf{x}_t|\hat{y}_c)$ (Eq. 1), will be high. Consider the two dimensional example illustrated in Fig. 7. In the figure, the ovals represent the covariance matrices that are characterized by the latent variables of the data points with the corresponding colors respectively. Namely, the ovals represent the covariance matrices that best explains the corresponding data points under the model. The two data points $\mathbf{x}_t$ and $\mathbf{x}_s$ show similar correlational structures to $\mathbf{x}_1$. Thus, $\mathbf{x}_t$ and $\mathbf{x}_s$ are well captured by the covariance matrix defined by $\hat{y}_1$, which is the latent variable for $\mathbf{x}_1$. On the other hand, covariance matrices which are optimized for data points with different correlational structures from those of $\mathbf{x}_t$ and $\mathbf{x}_s$ (for instance, $\mathbf{x}_2$ and $\mathbf{x}_3$ in Fig 7) return low conditional probability values of $\mathbf{x}_t$ and $\mathbf{x}_s$.

In image space, each axis would correspond to individual pixel values of images. Note that the representation achieves invariance to the pixel values of images as illustrated in Fig. 7) in that the model considers $\mathbf{x}_t$ to be more similar to $\mathbf{x}_1$ than $\mathbf{x}_2$ and $\mathbf{x}_3$ even though the two are closer to $\mathbf{x}_t$ in terms of the Euclidean distances based on pixel values. The analogy extends to the high dimensional space. The reason we do not use $p(\mathbf{x}|\hat{y}_t)$ to find similar data points to a target image $\mathbf{x}_t$ is that data points near the origin (for instance, $\mathbf{x}_o$ in Fig. 7) will be well captured by any multivariate Gaussian distributions regardless of their covariance information.

The second term in the metric, $p(\hat{y}_c)$, favors the correlational structures that can be described by sparse latent variables; in the case that two candidates return the same value of the conditional probabilities of the target image given their latent variables, the metric prefers the one that results in sparser representation, as it returns higher prior probability values (Eq. 3).

We demonstrate the usage of the joint probability described above using the image retrieval task; for a target image $\mathbf{x}_t$, we retrieve candidate scene images from a large scene

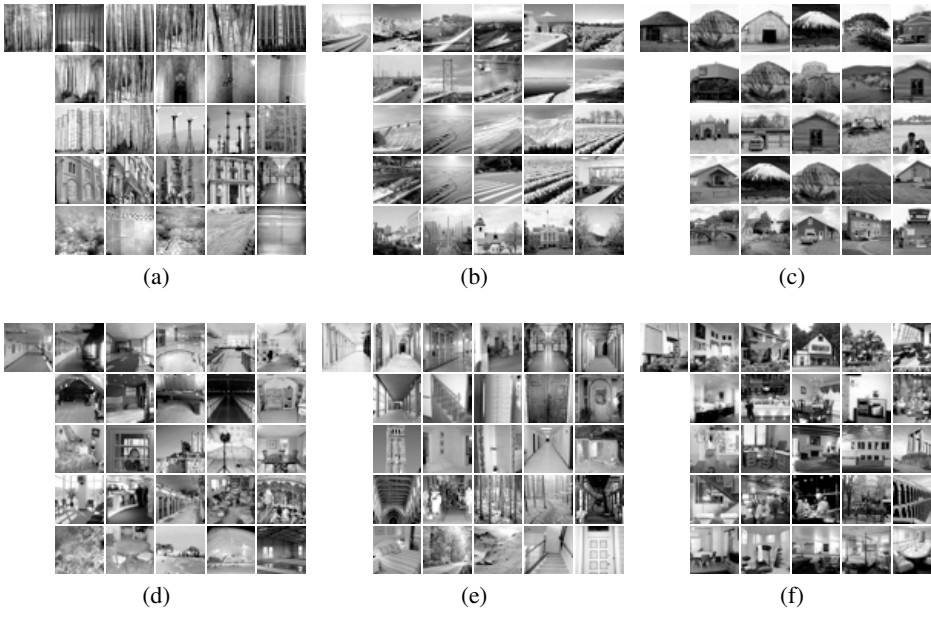


Figure 8: (a)–(h) Scene image retrieval results. The top left portion shows the target scene images. The retrieved images are ordered so that the left-most columns shows the most similar and the right-most columns show the 5 th similar candidate scene images to the targets. From the top to the bottom rows correspond to PCD, GIST(4×4), HOG(2×2), PHOG (3 levels), spatial pyramid of SIFT (3 levels). For (a)–(f) the target images are from the SUN database while the target images shown in (g)–(h) are not.

image pool (we used 108,754 images in the SUN database as the pool), whose latent variable returns the highest joint probability value with $\mathbf{x}_t$, or equivalently the lowest value of $-p(\mathbf{x}_t, \hat{y})$. We call this the probabilistic correlational distance (PCD) hereafter. In Fig. 8, we show the five most similar candidate scene images from 108,754 images retrieved with PCD, GIST, HOG, PHOG and spatial pyramid of SIFT. For GIST and HOG, we tried three different spatial scales (1×1, 2×2 and 4×4) and show the qualitatively best results. For all other representations than the distribution coding model, we used the Euclidean distances as similarity measures. Even though the model representation requires a small number of units to represent a scene image, the image retrieval results are qualitatively satisfactory. The distribution coding model achieves the efficiency by projecting scene images based on their characteristic features rather than representing scene images with fixed number of scales and orientations. In addition, it takes approximately 0.1 seconds to retrieve the similar images to targets using PCD which is fast enough for real-time image retrieval.

5 Quantitative evaluation of scene similarity measures

In this section, we quantitatively evaluate the similarity measure based on the distribution coding model on encoding the perceptual and the semantic similarities between scene images.

5.1 Perceptual similarities of scene images

To investigate whether the global correlational information encoded by the distribution coding model is consistent with the perceptual similarities between scene images, we conducted an experiment in which subjects were asked to select candidate scene images that were most similar to a target scene image in terms of spatial layout. In each trial, a target image from one of 397 semantic categories of the SUN database [23] was presented together with 25 randomly chosen *candidate* scene images. Subjects were allowed to select more than one candidate images if they were equally similar to the target images. We call the selected candidate images *similar* 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 of the scene, 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. Candidate images were chosen only from the same semantic categories as the target images, in order to control the difficulty of the tasks. Without such constraints, candidate images from different scene categories are too dissimilar to make meaningful judgements. In addition, using candidate images from the same category prevents subjects from depending on any semantic information to perform the task. Five subjects (one female; with normal or corrected to normal vision; 22-33 years old) participated in the experiment. We collected 2597 trials and the subjects selected 1.39 candidate images per trial on average (the number of candidate images selected per trial ranged from 0 to 13). Out of 2597 trials, subjects selected more than one similar images in 834 trials and selected zero similar images in 825 trials.

We evaluate the performances of various representations based on two criteria. The first one is the percentage of trials in which the similar images coincided with the closest candidate image to the target in a feature representation, the *closest* image. If a feature representation is consistent with the perceptual properties of images, the closest image will be perceived to be similar to the targets. The other criterion is the mean rank of the similar images when all the candidates in a trial are sorted in the ascending order in terms of the distances to the target in each representation. We assume that similar images will be more likely to have shorter distances to the targets than others and thus will have lower mean ranks of similar images if the distance is consistent with perception.

We use PCD introduced in the previous section as the scene similarity measure for the distribution coding model. For other representations, the Euclidean distances between the features extracted from images were adopted as the similarity measures. As we trained the distribution coding model and ICA on images of 32×32 resolutions, we downsampled the original images to 32×32 pixels and then extracted the corresponding features. For all other state-of-the-art representations, we extracted the features from images of 128×128 resolutions. As reported in Table 1, PCD shows the most consistencies with the perceptual experiment in terms of both criteria. Note that the

percentage criterion only takes into account the closest images whereas the mean ranks criteria considers all the similar images within a trial.

Table 1: Performance evaluation of various representations for the perceptual experiment on scene layout similarities. We show detailed performances for subcategories of scene images. IN, Out-Nat and Out-Man correspond to Indoor, Outdoor natural and outdoor manmade scenes, respectively.

Feature	Resolution	Percentage (%)				Mean Ranks			
		Total	IN	Out-Nat	Out-Man	Total	IN	Out-Nat	Out-Man
PCD	32×32	**19.3**	18.0	**22.5**	**17.2**	**7.01**	**5.91**	**7.16**	**6.93**
GIST(4×4)	128×128	16.3	16.0	17.8	15.2	10.7	9.67	10.4	11.0
ICA	32×32	10.4	8.00	12.7	8.92	11.7	10.7	11.1	12.2
HOG(2×2)	128×128	15.1	18.0	16.8	13.8	10.9	9.70	10.9	10.9
PHOG(L=3)	128×128	16.1	**20.0**	18.4	14.3	11.0	11.0	10.9	11.0
SIFT(L=3)	128×128	15.4	12.0	16.0	15.2	9.78	9.35	9.88	9.72

5.2 Perceptual spatial layouts

In the previous section, we discussed that the distance measure between scene images based on the density coding model predicted the similarity between scene images based on their spatial layouts. Here, we investigate the degree to which the latent representation based on the correlational structures of images is effective at predicting perceptual spatial layout properties.

Ross and Oliva [19] gathered ground truth human ratings of openness, mean depth and perspective in 1 to 6 scale (Figure 9). Openness of a scene refers to the quantity and location of boundaries in a scene. Openness 1 represents scenes with a large portion of unobstructed sky and dominant horizontal lines and openness 6 represents closed scenes. Mean depth refers to depth in a global sense related to the physical size of a scene. Scenes that are close by were rated with mean depth 1 and those which were far from the camera were rated with mean depth 6. Perspective means the degree of expansion in a scene which can be estimated by the angle between the camera and the perceptually dominant vanishing points in an image. Scenes with perpendicular camera angle to the vanishing points and thus have strong convergence between the parallel lines were rated with perspective 1 and scenes with surfaces at fairly uniform distances from the camera with perspective 6.

We investigate if the latent variables that encode the covariance information of scene images are predictive of the perceptual spatial layout ratings. We first infer the latent variables $\mathbf{y}$ (Eq.2) for the scene images with associated perceptual spatial layout ratings. Then, we fit linear regression functions for predicting individual perceptual ratings from the latent variables. We used 10-fold cross validation procedure for evaluating the test results. As the dataset consist of natural and urban scene categories, we fit the linear regression functions to separate categories and also to the total dataset (Table 2). We compare the latent variables $\mathbf{y}$ to GIST, HOG, PHOG, spatial pyramids and ICA coefficients. The latent variables $\mathbf{y}$ demonstrates superior performance to all other representations for predicting openness and depth ratings for natural and urban scenes

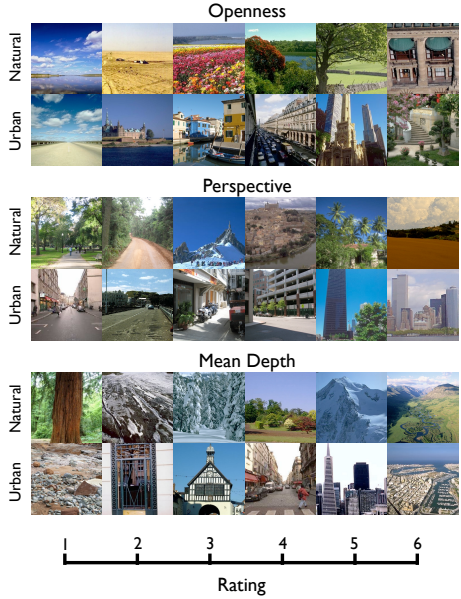


Figure 9: Representative scene images of the scene layout property ratings from [19]. See text for detailed explanation of the rating scale for each scene layout property.

each. For the perspective, which was reported to be hard to estimate for the subjects and also less consistent that depth and openness ratings, the latent variables $\mathbf{y}$ showed comparable prediction error to GIST, HOG and PHOG.

Table 2: RMSE of linear regression functions for predicting perceptual spatial layout ratings. *Urb* and *Nat* correspond to urban scenes and natural scenes, respectively.

Feature	Depth			Openness			Perspective		
	Urb	Nat	Total	Urb	Nat	Total	Urb	Nat	Total
$\mathbf{y}$ (32×32)	**0.59**	**0.70**	**0.66**	**0.80**	**0.92**	0.88	**1.20**	**1.18**	1.25
GIST (4×4, 128×128)	0.63	0.71	0.67	0.93	1.00	0.96	**1.22**	1.21	**1.21**
ICA (32×32)	0.71	0.84	0.77	1.21	1.32	1.23	1.46	1.28	1.36
HOG (4×4, 128×128)	0.65	0.75	0.72	1.02	1.08	1.08	1.23	**1.19**	1.29
PHOG(L=3, 128×128)	0.65	0.76	0.70	0.92	1.03	1.23	**1.21**	1.26	1.24
SIFT(L=3, 128×128)	0.80	0.82	0.71	0.91	0.97	**0.84**	1.50	1.60	1.34

6 Conclusion

We trained the distribution coding model to learn the correlational information on the whole scene images. The model parameters show global correlational structures reflecting the regularities found 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 significantly greater consistency with perceptual properties of the global structures in scene images. The probabilistic correlational distance can be used for image retrieval systems. Also the latent variable encoding the covariance information is significantly more predictive of perceptual spatial layouts (depth and openness) of scene images.

Our approach can be extended to larger size images for encoding more detailed local 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. As the model represents images or patches based on their adaptive representation rather than fixed number of scales and orientations, it could find match points more accurately especially in natural scenes in which points and lines are not defined by as high contrasts as indoor or manmade scenes. Extending the model training to images describing mainly of objects can also be useful for understanding object invariances under diverse viewing angles or nonrigid objects. Lastly, the analysis can be applied to face recognition system.

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. *CVPR*, pages 524–531, 2005.

[5] 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.

[6] 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.

[7] 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.

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

[9] 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.

[10] 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.

[11] 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.

[12] 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.

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

[14] 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.

[15] 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.

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

[17] H. Ouyang and A. Gray. Learning dissimilarities by ranking: from sdp to qp. In *Proceedings of the 25th international conference on Machine learning*, ICML '08, pages 728–735, New York, NY, USA, 2008. ACM.

[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] A. Torralba and A. Efros. Unbiased look at dataset bias. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[22] 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.

[23] 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.