Non-local Characterization of Scenery Images: Statistics, 3D Reasoning, and a Generative Model

Tamar Avraham and Michael Lindenbaum

Computer science department, Technion - I.I.T., Haifa 3200, Israel
{tammya,mic}@cs.technion.ac.il

Abstract. This work focuses on characterizing scenery images. We semantically divide the objects in natural landscape scenes into background and foreground and show that the shapes of the regions associated with these two types are statistically different. We then focus on the background regions. We study statistical properties such as size and shape, location and relative location, the characteristics of the boundary curves and the correlation of the properties to the region's semantic identity. Then we discuss the imaging process of a simplified 3D scene model and show how it explains the empirical observations. We further show that the observed properties suffice to characterize the gist of scenery images, propose a generative parametric graphical model, and use it to learn and generate semantic sketches of new images, which indeed look like those associated with natural scenery.

1 Introduction

> *By age 5 or 6 children develop a set of symbols to create a landscape that eventually becomes a single variation repeated endlessly. A blue line and sun at the top of the page and a green line at the bottom become symbolic representations of the sky and ground.* From: Drawing on the Right Side of the Brain. Betty Edwards, 1979 [1].

When we think of "scenery" or "natural landscape" images, we typically imagine a photograph or a painting, with a few horizontal background regions, each spanning the frame. The highest region would usually be the sky, while the lower regions might include mountains, trees, flowers, water (lake/sea), sand, or rocks. This work examines whether this intuition is justified, by analyzing image statistics and by modeling the 3D world and analyzing its 2D projections as imaged in typical scenery photography. We semantically divide the objects in natural landscape scenes into background and foreground and show that the shapes of the regions associated with these two types are statistically different. We then focus on the background regions. We study statistical properties such as size and shape, location and relative location, the characteristics of the boundary curves and the correlation of the properties to the region's semantic identity.

K. Daniilidis, P. Maragos, N. Paragios (Eds.): ECCV 2010, Part V, LNCS 6315, pp. 99–112, 2010.

These properties, which could be characterized as common world knowledge, have been used, in part, to enhance several computer vision algorithms. Nonetheless, they have not, to the best of our knowledge, been explicitly expressed, summarized, or computed.

This paper makes three contributions: First, we make several observations about image region properties, and collect empirical evidence supporting these observations from annotated segmentations of 2D scenery images (Section 2). Second, we discuss the imaging process of a simplified 3D scene model and show how it explains the empirical observations (Section 3). In particular, we use slope statistics inferred from topographic maps to show why land regions whose contour tangents in aerial images are statistically uniformly distributed appear with a strong horizontal bias in images taken from ground level. Third, we show that the observed properties suffice to characterize the gist of scenery images: In Section 4 we propose a generative parametric graphical model, and use it to learn and generate semantic sketches of new images, which indeed look like those associated with natural scenery. The novel characteristics analyzed in this work may improve many computer vision applications. In Section 5 we discuss our future intentions to utilize them for the improvement of top-down segmentation, region annotation, and scene categorization.

1.1 Related Work

Statistics of natural images play a major role in image processing and computer vision; they are used in setting up priors for automatic image enhancement and image analysis. Previous studies in image statistics (e.g., [2,3,4,5]) mostly characterized local low-level features. Such statistics are easy to collect as no knowledge about high-level semantics is required. Lately, computer-vision groups have put effort into collecting human annotations (e.g., [6,7,8]), mostly in order to obtain large ground-truth datasets that enable the enhancement and validation of computer vision algorithms. The availability of such annotations enables the inference of statistics on semantic characteristics of images. A first step was presented in [6] where statistics on characteristics of human segmentation were collected. In [7] a few interesting statistics were presented, but they mainly characterize the way humans segment and annotate. We follow this direction relying on human annotations to suggest characteristics that quantify high-level semantics.

The importance of context in scene analysis was demonstrated a while ago [9] and used intensively in recent years for improving object detection by means of cues pertaining to spatial relations and co-occurrence of objects [10,11,12], annotated segments [13], or low-level scene characteristics and objects [14,15]. Part of the work presented here focuses on the co-occurrence of and spatial relations between background objects. Objects are semantically divided into background and foreground, implying that image analysis applications may benefit from treating objects of these two classes differently. This is related to the *stuff* vs. *things* concept [16].

We found that the background-region boundary characteristics correlate with the identity of the lower region. This observation is consistent with the

observation made in figure-ground assignment studies, that of the two regions meeting in a curve, the lower region is more likely to be the "figure", i.e., of lower depth [17,18]. Our work is also related to the recent successful approach which uses learning to model the relation between image properties and the corresponding 3D structure (e.g., [19,20,21]). The approach in [21], for example, associates the images with particular classes of geometrically specified 3D structures. We focus here on the wide class of scenery images with a variety of semantically specified regions, and provide an image model (supported by 3D scene analysis) characterizing the large scale image properties.

2 Observations and Evidence

In this section we present some observations on the appearance of background regions in landscape images. After showing how the statistics of their general shape differ from those of foreground objects, we discuss their relative location and characteristics of their contours.

We use fully annotated landscape images included in the Labelme toolbox [7]. Of the images used in [22], where outdoor images were divided into eight categories, we used all the images of the three natural landscape categories: *coast*, *mountain*, and *open country* (for a total of 1144 256X256 images).

With the Labelme toolbox, a Web user marks polygons in the image and freely provides a textual annotation for each. This freedom encourages the use of synonyms and spelling mistakes. Following [7], synonyms were grouped together and spelling mistakes were corrected.(For details see [23].)

2.1 The General Shape of Background vs. Foreground Objects

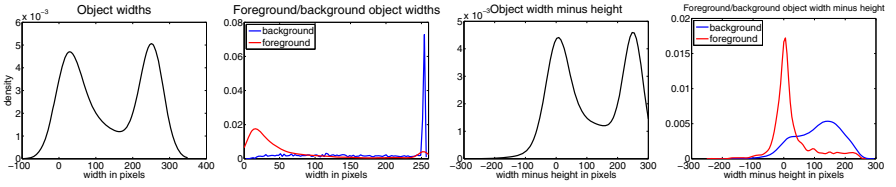
We semantically divide the annotated objects into two sets: *background objects* and *foreground objects*. The background set includes all objects belonging to the following list: *sky*, *mountain*, *sea*, *trees*, *field*, *river*, *sand*, *ground*, *grass*, *land*, *rocks*, *plants*, *snow*, *plain*, *valley*, *bank*, *fog bank*, *desert*, *lake*, *beach*, *cliff*, *floor*. Foreground objects are defined as those whose annotation does not belong to that list. (Note that while *trees*, *rocks*, and *plants* are considered *background objects*, *tree*, *rock*, and *plant* are considered *foreground objects*.) For a summary of the occurrence of each of the background and foreground labels see [23]. Differences in the distribution of the size and aspect ratios of the bounding box of these two classes give rise to the following observations:

Observation 1: Many background objects exceed the image width.

The background objects are often only partially captured in the image. See Fig. 1(a) and Fig. 1(b) for background vs. foreground object width statistics. Note the sharp bimodality of the distribution.

Observation 2: The background objects are wide and of low height while foreground objects' shape tend to be isotropic.

Although the entire background object width is usually not captured, the height of its annotated polygon is usually small relative to the height of the image. See



(a) (b) (c) (d)

Fig. 1. Bounding boxes of imaged background objects are usually low height horizontal rectangles, while those of imaged foreground objects tend to be squares: (a) width density estimation (by kernel smoothing) of annotated objects from the Labelme dataset (where width is the difference in pixel units between the rightmost and the leftmost points in the annotated polygon. All images are 256×256); (b) width density estimation of background and foreground objects taken separately; (c) width minus height density estimation of annotated objects; (d) width minus height density estimation of background and foreground objects taken separately. The distributions in (a),(c) were generated by an equal number of foreground and background objects. A random subset of background objects was used to compensate for the larger number of background objects in this dataset.

Fig. 1(c) and Fig. 1(d): the width and height difference of foreground objects is distributed normally with zero mean, while the width and height difference of background objects significantly favors width. This implies that bounding boxes of imaged background objects are low height horizontal rectangles, while bounding boxes of imaged foreground objects tend to be squares. (Note that all the images in this dataset are squares, so the horizontal bias is not due to the image dimensions.) See further analysis and discussion on the horizontalness of background regions in Section 3.

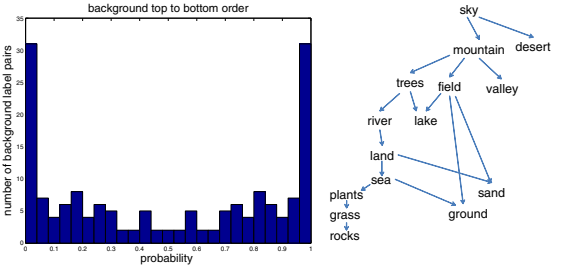
2.2 The Top-Down Order of Background Objects

Because background objects tend to be wide—frequently spanning the image horizontally though not vertically—each landscape image usually includes a few background regions, most often appearing one on top of the other.

Observation 3: The relative locations of types of background are often highly predictable.

It is often easy to guess which background type will appear above another. For instance, if an (unseen) image includes sky and ground, we know that the sky will appear above the ground. Here we extend this ostensibly trivial "sky is above" observation and test above-and-below relations for various pairs of background types. Let $\mathcal{I}$ denote the set of all landscape images. Let $A - B$ denote that background type A appears above background type B in image $I \in \mathcal{I}$ (e.g., A=trees, B = mountain). We estimate the probability for A to appear above B (or B to appear above A), given that we know both appear in an image:

$$p_{A-B} = p(A - B | A, B \in I) \simeq \frac{|\{I \in \mathcal{I} | A, B \in I, A - B\}|}{|\{I \in \mathcal{I} | A, B \in I\}|} \quad . \tag{1}$$



(a) (b)

Fig. 2. Expected relative location of background regions. For most background type pairs, there is a strong preference for one type to appear above the other. (a) The probability for a background region of identity A to appear above a background region with identity B, summarized in a histogram for various background identity pairs. (b) Topological ordering of background identities can be defined: this DAG (Directed Acyclic Graph) is associated with the reachability relation $R : \{(A, B)|p_{A-B} > 0.7\}$.

See Fig. 2 for a histogram of p_{A-B} for A and B being two background identities, $A \neq B$. The histogram is symmetric as $p_{A-B} + p_{B-A} = 1$. There are 22 background categories. Out of 231 possible pairs, only 116 appear at least once in the same image. The histograms consider only pairs that coappeared at least 5 times (83 pairs). Most pairs show a clear preference for the more likely relative location. The most obvious is sky, which appears above all other background categories. However, some examples for pairs for which $p_{A-B} > 0.9$ are mountain-lake, trees-plants, mountain-beach, trees-rocks, plain-trees. For 84% of the pairs, $\max(p_{A-B}, p_{B-A}) > 0.7$. The dominant order relations induce a (partial) topological ordering of the background identities and can be described by a DAG (Directed Acyclic Graph). The DAG in Fig. 2(b) is associated with the reachability relation $R : \{(A, B)|p_{A-B} > 0.7\}$, i.e., there is a directed path in the graph from A to B if and only if A appears above B in more than 70% of the images in which they coappear. As evident here, learning the typical relative locations of background regions is informative.

2.3 Contours Separating Background Regions

If a background region A appears above a background region B, it usually means that B is closer to the photographer and is partly occluding A [17,18]. Hence, we can say that a contour separating background regions is usually the projection of the closer (lower) background region's silhouette, and usually has characteristics that can be associated with this background type.

Observation 4: The characteristics of a contour separating two background regions correlates with the lower region's identity.

Consider Fig. 3. The curves in Fig. 3(b-d) are associated with the background object classes 'mountain', 'trees', and 'grass', respectively. See also Fig. 3(e)-(g), for contours associated with different background objects. When the lower

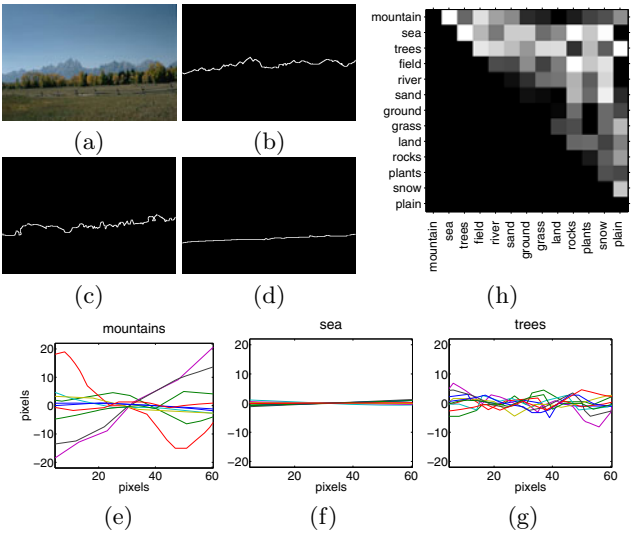


Fig. 3. Characteristics of background region boundaries. (a)-(d) An image and its hand segmentation [6]. (e)-(g) A few sample contour segments associated with background object classes 'mountain', 'sea', and 'trees' (from Labelme). (h) Classification accuracies for two-class background identity, based only on the appearance of the region's upper boundary. The accuracy of classification is displayed in linear gray scale (black for all values below 0.5 and white for all values above 0.8).

background object is of type sea, grass or field, the boundary is usually smooth and horizontal, resembling a DC signal. For background objects such as trees and plants, the boundary can be considered as a high frequency 1D signal. For background objects of type 'mountain', the associated boundaries usually resemble 1D signals of rather low frequency and high amplitude.

Adopting a signal representation, we checked how informative the contours are for discriminating between background identities: The Labelme landscape images were randomly divided into equally sized training and validation sets. For each background labeled region, the upper part of its contour was extracted and cut to chunks of 64-pixels length. Each chunk was FFT transformed, and the norms of 32 coefficients (2-33) were added to a training or validation set associated with the background label. Only labels for which the training set included at least 30 'signals' were further considered. For each pair of labels we checked whether the two associated classes could be distinguished using ONLY the upper part of their boundary. We used an SVM classifier with an RBF kernel. (To avoid bias due to different class size, which would have made discrimination easier, we took equal sized training sets and validation sets from the two classes, by randomly selecting some members of the larger set.) Fig. 3(h) summarizes the accuracies of the two-class classifiers. While the information in the contour's shape cue discriminates well between several pairs of background types, it cannot discriminate between all pairs. Better results may be obtained by adding local

properties as discussed in Section 5. In Section 4 we show that the contour shape information together with the relative location may be used to specify a generative model that captures the gist of scenery images.

3 Why Are Background Regions Horizontal? A 3D Analysis

In Section 2 we have statistically shown that imaged land region boundaries have a strong horizontal bias. To account for this empirical finding, we model the 3D world and analyze its 2D projections imaged in typical scenery photography. We start by a simplified 'flatland' model, continue by considering also land coverage (e.g., vegetation), and finally terrain elevation. In all these cases, we show why land regions whose contour tangents in aerial images are uniformly distributed appear with a strong horizontal bias in scenery filmed on the ground.

3.1 Flatland

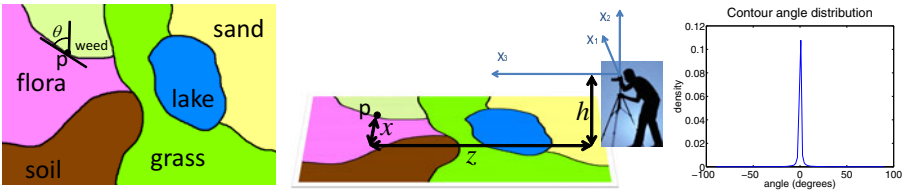
> *Place a penny on the middle of one of your tables in Space ... look down upon it. It will appear a circle....gradually lower your eyes ... and you will find the penny becoming more and more oval to your view....* From Flatland, by Edwin A. Abbott, 1884 [24].

We first consider a natural terrain in a flat world with no mountains, no valleys, and vegetation of zero height. This terrain may be divided into a few regions, each with different "clothing", as depicted from an aerial view in Fig. 4(a). Consider the contour lines dividing the different regions. Let Θ be the set of tangent angles for all such contours, measured relative to some arbitrary 2D axis on the surface. It is reasonable to assume that the angles in Θ are uniformly distributed in the range $[0°, 360°)$. Now consider a person standing on that surface at an arbitrary point, taking a picture. Let Θ' be the set of angles that are the projections of the angles in Θ on the camera's image plane. How is Θ' distributed?

For simplicity, we adopt the pinhole camera model. Let a point p on a contour line be located at $(x, -h, z)$ in a 3D orthogonal coordinate system originating at the camera pinhole. (See Fig. 4(b).) Redefine θ as the angle of the tangent to the contour line at p, relative to the 1st axis. The angle θ', associated with the projected contour on the camera's sensor is

$$\tan\theta' = \frac{h\tan\theta}{z - x\tan\theta} \quad . \tag{2}$$

For details see [23]. See Fig. 4(c) for a plot of the distribution of Θ'. The strong peak around 0° explains why background regions tend to be wide and horizontal in scenery images (as statistically shown in Section 2.1).



(a) (b) (c)

Fig. 4. The tangents of background object contours when imaging a flat terrain. (a) A schematic illustration of an aerial image; (b) a view from height h. A point p that lies on a land region boundary is located at $(x, -h, z)$ relative to a 3D orthogonal coordinate system originating at the camera pinhole; (c) the distribution of the tangent of boundary lines in such an image, assuming that the tangents of aerial image boundaries are uniformly distributed, $\theta \sim U[0, 180)$, $h = 2$[m], $z \sim U[0$[m], 1000[m]], and $x \sim U[0[m], 500[m]]$.

3.2 Land Cover

Now we extend the flatland model and consider a flat terrain with protruding coverage, e.g., sand, gravel, or rock covered regions, fields of flowers, or even forests. Each such region's cover is often of approximately equal height. Then, the profile of this land (slicing through any vertical plane) can be considered as a piecewise constant function.

Consider again the photographer at an arbitrary point on the flat terrain. First consider the case where the cover is lower than the camera (e.g., bushes, pebbles). The cover of a raised region would occlude part of the more distant regions. The distribution of angles associated with the tangents of imaged contours describing the upper contour of such cover is even more concentrated near the origin, as the height difference of points on the cover and the pinhole is smaller compared to the height difference in flatland. When the land cover is higher than the camera (e.g., forest), the region cannot be viewed from above and only the side facing the camera will be captured. Angles on the upper contour, at height H project to image angles, θ', where $\tan\theta' = \frac{(H-h)\tan\theta}{z - x\tan\theta}$. Typically, trees are only a few meters high while the viewing distance z for landscape images is usually much larger. Therefore, the statistical shortening of the contour angles still holds.

Naturally, the land cover height is not constant, but characterized by some nominal value with small local perturbation. These perturbations may significantly change the local projected angle but not the general direction of the contour, which stays close to horizontal.

3.3 The World Is Wrinkled: Ground Elevation and Slope Statistics

Obviously, the earth's terrain is not flat. Its surface is wrinkled by mountains, hills and valleys. To approximately express how ground elevations affect the appearance of background object contours in images, we rely on a slope statistics database [25]. This dataset includes histograms over 8 variable size bins for each

land region of about 9 square kilometers. (The bins are nonuniform and are specified between the slope limits of 0%, 0.5%, 2%, 5%, 10%,15%, 30%, 45%, and infinity.) We use the average histogram. To get a distribution, we approximate the slope distribution within the bin as uniform. See Fig. 5(b). We also use a histogram of the maximum slope over the land regions (Fig. 5(c)).

The slope statistics affect two landscape image contour types: (1) The contours of mountains associated with occluding boundaries (e.g., skylines). (2) The contours between different types of regions on the terrain.

The distribution depicted in Fig. 5(c) provides a loose upper bound for the expected distribution of projected tangent angles associated with the former set. Even so, the horizontal bias is apparent.

To account for the effect of ground elevation on the latter type of background contours, we extend the analysis suggested in Section 3.1. Instead of considering an angle θ lying on a flat terrain, we consider θ to lie on an elevated plane with slope gradient angle φ. See Fig. 5(a). The plane is rotated relative to the image plane, forming an angle ω with the X_1 axis. The point p is at height H relative to the camera height. The projected tangent angle θ' is given by

$$\tan\theta' = \frac{H(cos\theta \sin\omega + \sin\theta\cos\varphi\cos\omega) - z\sin\theta\sin\varphi}{x(cos\theta\sin\omega + \sin\theta\cos\varphi\cos\omega) - z(\cos\theta\cos\omega - \sin\theta\cos\varphi\sin\omega)} \quad (3)$$

For details see [23]. To get an idea how θ' is distributed, we make several reasonable assumptions: θ is uniformly distributed as before, φ is distributed as the slope angle distribution (Fig. 5(b)), and ω is uniformly distributed $U[-90°, 90°]$. The distribution of H was estimated by sampling an elevation map [25], using the height difference between pairs of locations up to 9km apart. See an analytic plot of the distribution of θ' in Fig. 5(d).

The above analysis isn't perfect from either a geometrical, a topographical, or an ecological point of view; e.g., we do not account for the roundness of the world, we assume that the camera is levelled with the ground, we assume independency between the slope steepness and the imaging height difference, and we do not consider dependencies between the steepness of slopes and the location of different land regions. For instance, the slope of a lake is always zero. Nevertheless, we believe the horizontal bias of background contours, as observed empirically, is sufficiently accounted for by the simplified analysis described here.

4 A Generative Model

The observations and the statistical quantification presented in Section 2 enable us to propose the following generative model for scenery image sketches. See, e.g., Fig. 3(a)-(d). Our model considers the top-down order of background regions, the relative area covered by each, and the characteristics of their boundaries, and assigns a probability for each possible annotation sequence.

Let $S = (h_1, ...h_n, S_1, S_2, ..., S_{n-1})$ be the description of a background segmentation for an image divided into n background segments, ordered from the highest in the image ($i = 1$) to the lowest ($i = n$). h_i is the mean height of region

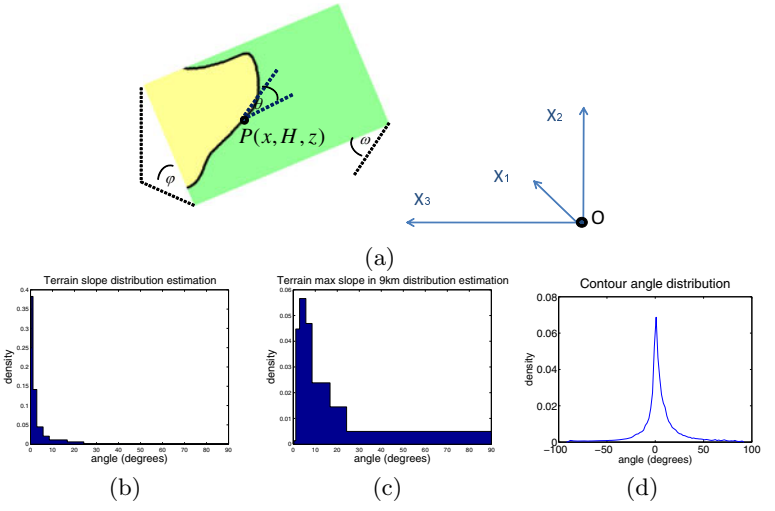


Fig. 5. Distribution of background object contour angles in a "wrinkled" world. (a) A point p lies on a boundary between land regions. It is located on an elevated slope with gradient angle φ. The infinitesimal plane is rotated at an angle ω relative to X_1 axis. (b) Estimated terrain slope distribution using the IIASA-LUC dataset [25]. (c) Estimated distribution of the maximum slope over land regions, each covering approximately 9 square kilometers. (d) The distribution of the tangents of imaged boundaries, following the analysis in the text.

i, $\sum_i h_i = 0$. S_i describes a '1-D signal' associated with the boundary between i and $i+1$. Let $\boldsymbol{l} = (l_1, ..., l_n)$ be a labeling sequence, where $l_i \in L$, and L is the set of background labels. We shall use the approximated distribution

$$P(\boldsymbol{l}|S) = \frac{P(S|\boldsymbol{l})P(\boldsymbol{l})}{\sum_{\boldsymbol{l}\in L^n} P(S|\boldsymbol{l})P(\boldsymbol{l})} \propto P_1(\boldsymbol{l}) \prod_{i=1,...,n} P_2(h_i|l_i) \prod_{i=1,...,n-1} P_3(S_i|l_{i+1}) \quad . \tag{4}$$

This approximation assumes that the height of a region depends only on its identity, and that a boundary's characteristics depend only on the identity of the corresponding lower region. Other dependencies are ignored.

The next three sections discuss the distributions P_1, P_2 and P_3.

4.1 A Markov Network for Modeling the Top-Down Label Order

P_1 is the probability of a scenery image annotation. We use a Markov network to represent possible label sequences. Let $m = |L|$ be the number of possible background labels. The network has $m + 2$ nodes (statuses). The first m are associated with the members of L. In addition, there is a starting status denoted 'top' and a sink status denoted 'bottom'. Let M be the transition matrix. $M(l_i, l_j)$ is the probability to move from status associated with label l_i to status associated with l_j, i.e., that a region labeled l_i appears above a region labeled

l_j in an image. $M(\text{top}, l_i)$ and $M(l_i, \text{bottom})$ are the probabilities that a region with label l_i is at the top/bottom of an image, respectively. Then:

$$P_1(\boldsymbol{l} = (l_1, ..., l_n)) = M(\text{top}, l_1) \prod_{i=1,...,n-1} M(l_i, l_{i+1}) \; M(l_n, \text{bottom}) \,, \qquad (5)$$

i.e., the probability for a labeling sequence $(l_1, ..., l_n)$ is equal to the probability for a random walk starting at the initial network state to go through the states corresponding with $(l_1, ..., l_n)$, in that order, and to then continue to the sink status. We use a dataset of images for which the sequences of background labeling are known, e.g., the Labelme landscape images, and set the parameters of the model (i.e., the matrix M) by counting the occurrences of the different 'moves'.

4.2 A Normal Distribution for the Height Covered by Each Region

P_2 models the distribution of the relative height of an image region associated with a certain label. Here we simply use a normal distribution, learning the mean and variance of each region type's relative height.

4.3 Modeling Background Contours with PCA

$P_3(S_i|l_k)$ is the probability of a contour with appearance S_i to separate two background regions, the lower being of type l_k. In Section 2.2 we have shown that S_i and l_k are correlated (observation 4). To estimate the probability from examples we use PCA approximation (Principle Component Analysis [26]). Given a training set of separation lines associated with background type l_k, each separation line is cut to chunks of 64-pixel length[1]. Each chunk's mean value is subtracted, and PCA is performed, resulting in the mean vector $\bar{\mu}$, the first κ principle components Φ (a 64xκ matrix) and the corresponding eigen values $\bar{\lambda} = (\lambda_1, ..., \lambda_\kappa)$. κ is chosen so that 95% of the variation in the training set is modeled.

The PCA modeling allows both computation of the probability of a new separating line S_i, cut to chunks $S_{i1}, ..., S_{im}$, to belong to the learned distribution Ω, and generation of new separating lines belonging to the estimated distribution.

4.4 Generative Model Demonstration

We can now use this model to generate sketches of new images. To generate a sketch, a sequence of labels is first drawn from a random walk on the transition matrix (P_1). Then, heights are randomly picked from the normal distributions (P_2) and normalized. Finally, for each separating line indexed i, four 64-length chunks $S_{i,1}, ..., S_{i,4}$ are generated by $G = \bar{\mu} + \bar{b} \cdot \Phi$, where $b_j \sim N(0, \sqrt{\lambda_j})$, $j = 1, ..., \kappa$. (Each chunk is generated independently, ignoring appearance dependencies between chunks of the same line.) The leftmost point of chunk $S_{i,1}$

[1] Cutting the 'signals' into chunks also allows us to use separating lines from the training set that do not horizontally span the entire image or that are partly occluded by foreground objects. Moreover, it enlarges the training set, by obtaining a few training items (up to 4 chunks) from each separating contour.

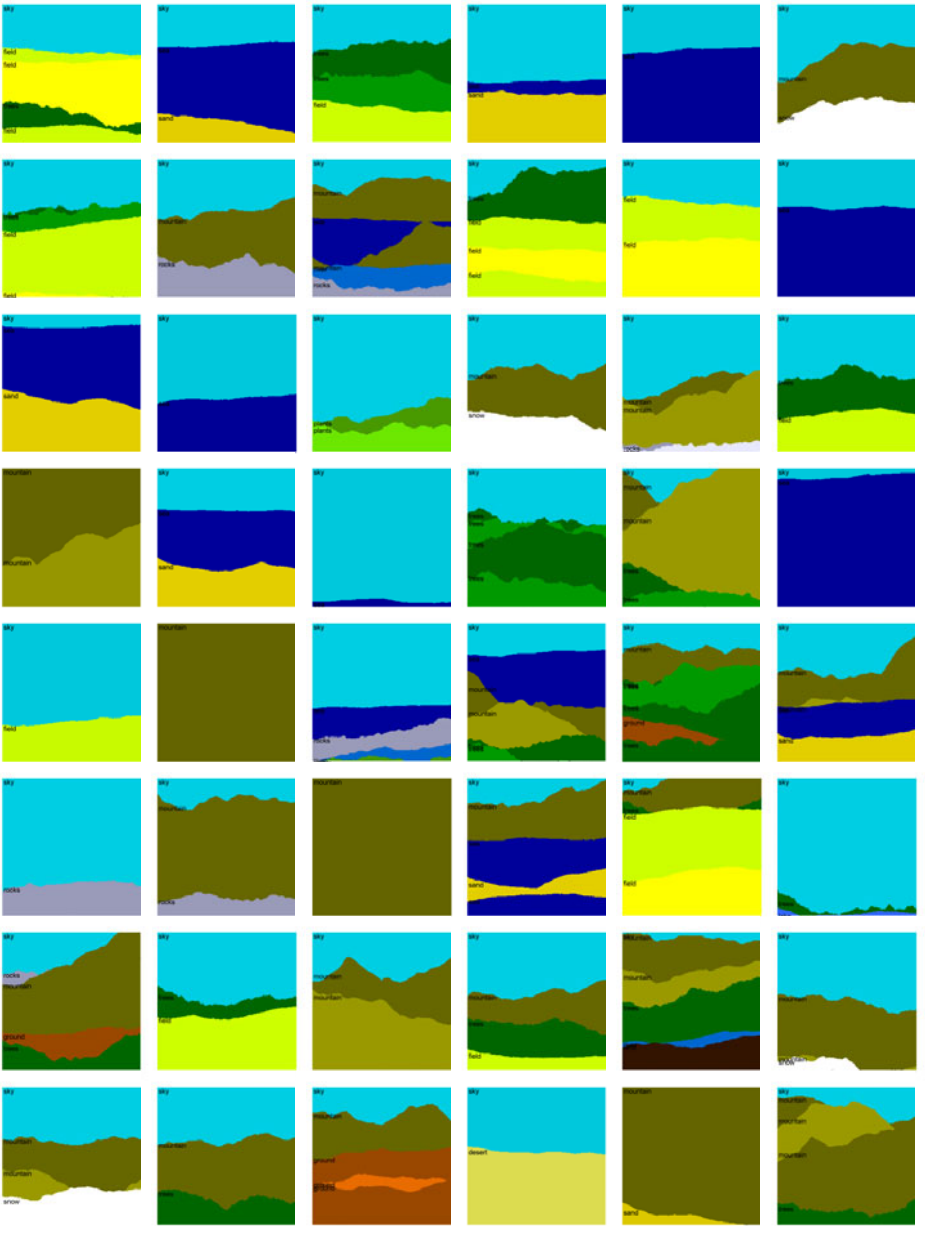


Fig. 6. A random sample of 'annotated' landscape images generated by our model. The regions are colored with colors associated with their annotation (sky regions are colored in blue, ground regions are colored in brown, etc.) Best viewed on a color computer screen.

is placed at image coordinate $(0, \sum_{j=1}^{i} h_j)$ (where $(0,0)$ is the top-left image corner). Chunk's $S_{i,m}$ $(m = 2,3,4)$ leftmost point is connected to the rightmost part of $S_{i,m-1}$. See Fig. 6 for a random sample of 'annotated' scenery landscape image sketches generated by the model.

To evaluate the generated, annotated images, we took a random sample of 50 and asked two participants naive to this research, aged 7 and 37, to say whether they seem to be the annotations of real landscape photos. The first participant answered 'yes' for 37 images, 'no' for 5, and was not sure about 8. The second participant answered 'yes' for 44 images, 'no' for 3 and 'not sure' for 3.

5 Discussion

This work focused on characterizing scenery images. Intuitive observations regarding the statistics of co-occurrence, relative location, and shape of background regions were explicitly quantified and modeled, and 3D reasoning for the bias to horizontalness was provided.

Our focus was on non-local properties. The generated image sketches, which seem to represent a wide variety of realistic images, suggest that the gist of such images is described by those properties. The proposed model provides a prior on scenery image annotation. In future work we intend to integrate local descriptors (see e.g. [27,28]) into our model and to then apply automatic annotation of segmented images. The large scale model introduced here should complement the local information and lead to better annotation and scene categorization [27]. Relating the contour characteristics to object identity can be useful for top-down segmentation (e.g., [13]). Specifically, it may address the "shrinking bias" of graph-cut-based methods [29].

A more complete model of scenery images may augment the proposed background model with foreground objects. Such objects may be modeled by location, size, shape, and their dependency in the corresponding properties of other co-occurring foreground objects and of the corresponding background regions.

One immediate application would be to use the probabilistic model to automatically align scenery pictures, similar to the existing tools for automatic alignment of scanned text. Some would find an artistic interest in the generated scenery sketches themselves, or may use them as a first step to rendering.

References

1. Edwards, B.: Drawing on the Right Side of the Brain. Tarcher Publishing (1979)
2. Lee, A.B., Mumford, D., Huang, J.: Occlusion models for natural images: A statistical study of a scale-invariant dead leaves model. IJCV 41, 35–59 (2001)
3. Srivastava, A., Lee, A., Simoncelli, E.: On advances in statistical modeling of natural images. Journal of Mathematical Imaging and Vision 18, 17–33 (2003)
4. Heiler, M., Schnörr, C.: Natural image statistics for natural image segmentation. IJCV 63, 5–19 (2005)
5. Ruderman, D.: The statistics of natural images. Computation in Neural Systems 5, 517–548 (1994)

6. Martin, D., Fowlkes, C., Tal, D., Malik, J.: A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: ICCV (2001)
7. Russell, B., Torralba, A.: Labelme: a database and web-based tool for image annotation. IJCV 77, 157–173 (2008)
8. Bileschi, S.: StreetScenes: Towards scene understanding in still images. Ph.D. Thesis, EECS, MIT (2006)
9. Rimey, R.D., Brown, C.M.: Control of selective perception using bayes nets and decision theory. IJCV 12, 173–207 (1994)
10. Desai, C., Ramanan, D., Fowlkes, C.: Discriminative models for multi-class object layout. ICCV (2009)
11. Rabinovich, A., Vedaldi, A., Galleguillos, C., Wiewiora, E., Belongie, S.: Objects in context. ICCV (2007)
12. Gupta, A., Davis, L.S.: Beyond nouns: Exploiting prepositions and comparative adjectives for learning visual classifiers. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part I. LNCS, vol. 5302, pp. 16–29. Springer, Heidelberg (2008)
13. He, X., Zemel, R.S., Ray, D.: Learning and incorporating top-down cues in image segmentation. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3951, pp. 338–351. Springer, Heidelberg (2006)
14. Torralba, A.B.: Contextual priming for object detection. IJCV 53, 169–191 (2003)
15. Russell, B.C., Torralba, A.B., Liu, C., Fergus, R., Freeman, W.T.: Object recognition by scene alignment. NIPS (2007)
16. Heitz, G., Koller, D.: Learning spatial context: Using stuff to find things. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part I. LNCS, vol. 5302, pp. 30–43. Springer, Heidelberg (2008)
17. Vecera, S., Vogel, E., Woodman, G.: Lower region: A new cue for figure-ground assignment. Journal of Experimental Psychology: General 131, 194–205 (2002)
18. Fowlkes, C., Martin, D., Malik, J.: Local figureground cues are valid for natural images. Journal of Vision 7, 1–9 (2007)
19. Torralba, A.B., Oliva, A.: Depth estimation from image structure. IEEE T-PAMI 24, 1226–1238 (2002)
20. Hoiem, D., Efros, A.A., Hebert, M.: Recovering surface layout from an image. IJCV 75, 151–172 (2007)
21. Nedović, V., Smeulders, A., Redert, A.: Depth information by stage classification. In: ICCV (2007)
22. Oliva, A., Torralba, A.: Modeling the shape of the scene: a holistic representation of the spatial envelope. IJCV 42, 145–175 (2001)
23. (Supplimentary material)
24. Abbott, E.A.: Flatland: A Romance of Many Dimensions (1884)
25. Fischer, G., Nachtergaele, F., Prieler, S., van Velthuizen, H., Verelst, L., Wiberg, D.: Global Agro-ecological Zones Assessment for Agriculture (GAEZ 2007). IIASA, Laxenburg, Austria and FAO, Rome, Italy (2007)
26. Lanitis, A., Taylor, C.J., Cootes, T.F.: Automatic interpretation and coding of face images using flexible models. IEEE-PAMI 19, 743–756 (1997)
27. Vogel, J., Schiele, B.: Semantic modeling of natural scenes for content-based image retrieval. IJCV 72, 133–157 (2007)
28. van de Sande, K., Gevers, T., Snoek, C.: Evaluating color descriptors for object and scene recognition. IEEE T-PAMI 99 (2009)
29. Vicente, S., Kolmogorov, V., Rother, C.: Graph cut based image segmentation with connectivity priors. In: CVPR (2008)