Consistent Temporal Variations in Many Outdoor Scenes

Nathan Jacobs, Nathaniel Roman, and Robert Pless
Department of Computer Science and Engineering
Washington University in St. Louis St. Louis, MO, 63117
{jacobsn,ngr1,pless}@cse.wustl.edu

Abstract

This paper details an empirical study of large image sets taken by static cameras. These images have consistent correlations over the entire image and over time scales of days to months. Simple second-order statistics of such image sets show vastly more structure than exists in generic natural images or video from moving cameras. Using a slight variant to PCA, we can decompose all cameras into comparable components and annotate images with respect to surface orientation, weather, and seasonal change. Experiments are based on a data set from 538 cameras across the United States which have collected more than 17 million images over the the last 6 months.

1. Introduction

What can we learn from a static camera that observes the same environment over long time periods? The statistics of image variations observed from such cameras has not been well studied, despite the fact that an enormous number of fixed cameras are capturing images every minute. Here we characterize patterns of variation common to natural sequences from any static camera. Our study is based on a data set of images taken every half hour over the last 6 months from 538 cameras distributed across the United States.

We initially follow the methods and approach of work characterizing the statistics of arbitrary natural image patches and windows of short video clips. But for video taken from a single viewpoint, the same analytic tools find much more specific statistical correlations. These correlations relate to important scene features. For example, image regions that share geometric features such as surface normal and depth have correlated responses to lighting changes. Clustering of appearance changes [4] and explicit modeling of the physics of scattering media [5] have shown impressive results on segmenting scene structure and weather patterns of long sequences of images from a static camera [6]. We claim that these structures are available in data from static cameras without complicated algorithms or physical modeling, using only principal component analysis over time scales of days, weeks, and months. Furthermore, static cameras show surprisingly similar types of variation which can be unified into a canonical decomposition. This supports the automatic annotation, in any static camera, of the scene structure at a pixel location. Figure 1 shows an example of this automated annotation for data taken for a month from 12 cameras with colors indicating sky (light blue), trees (light green), eastward facing wall (orange),

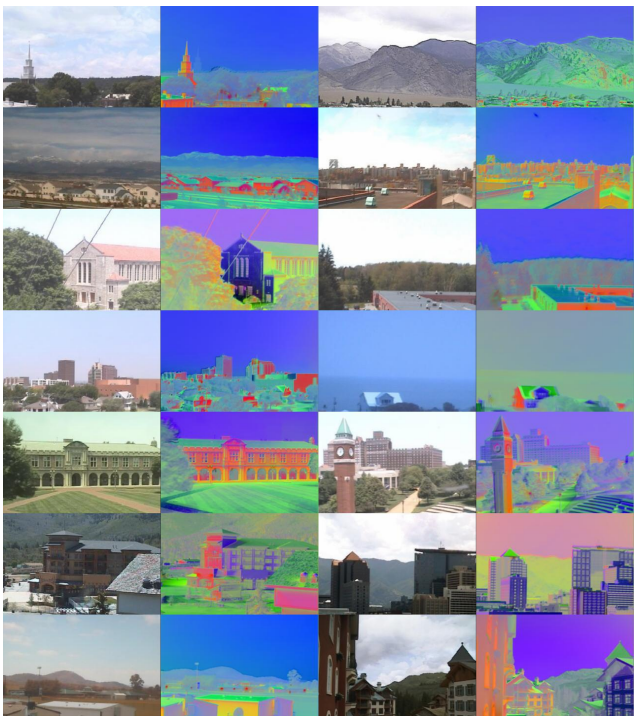


Figure 1. The components of the canonical day decomposition code for lighting variations. The above shows a collection of pairs of an example image from a camera, and a false color image made from the first 3 components of the canonical day decomposition. The colors indicate sky (light blue), trees (light green), eastward facing wall (orange), westward facing wall (blue).

westward facing wall (blue).

1.1. Background and Related Work

Studies of natural images have considered second-order statistics through the PCA decomposition [3], and, more recently, using higher order statistics and Independent Components Analysis (ICA). When ICA is applied to natural image patches to find optimal sparse codes, it gives basis images that appear very similar to receptive fields in the visual cortex (for example [8, 10]). However, these statistics are only computed for relatively small patch sizes because in natural images there are only weak correlations between pixels that are far apart.

In replicating these studies on image patches taken from the same location in a static camera, we find empirical evidence (see Section 3) that location specific bases are much more informative than patch bases developed on generic natural image patches. These bases reflect stronger correlations between distant pixels, and these strong correlations exist over extremely large patches.

Correlations between multiple images have been studied for natural video where the dominant cause of image change is camera motion. In this case, small space time patches have non-separable spatial and temporal correlations [2], and in optimal sparse codes designed for such time-varying natural imagery, nearly all of the basis functions code for motion [7]. These are also studied largely on small patches as correlation between pixels decrease with longer spatial and temporal distances.

In Section 4 we explore temporal correlations in natural video sequences at time scales of a day and longer. We find strong correlations that exist in every static camera across the entire image and throughout the entire 6 month length of our data set.

2. AMOS: Archive of Many Outdoor Scenes

The AMOS dataset[1] consists of over 17 million images captured since March 2006 from 538 outdoor webcams. This dataset is unique in that it contains significantly more scenes than in previous datasets [6] of natural images from static cameras. This enables us to empirically answer question that were previously untenable.

The cameras in the dataset were selected by a group of graduate and undergraduate students using a standard web search engine. Images from each camera are captured several times per hour using a custom web crawler that ignores duplicate images and records the capture time. The images from all cameras are 24-bit JPEG files that vary in size from 316×240 to 2048×1536, with the majority being 320×240.

[1]The dataset is available to the community at http://www.cse.wustl.edu/amos/.

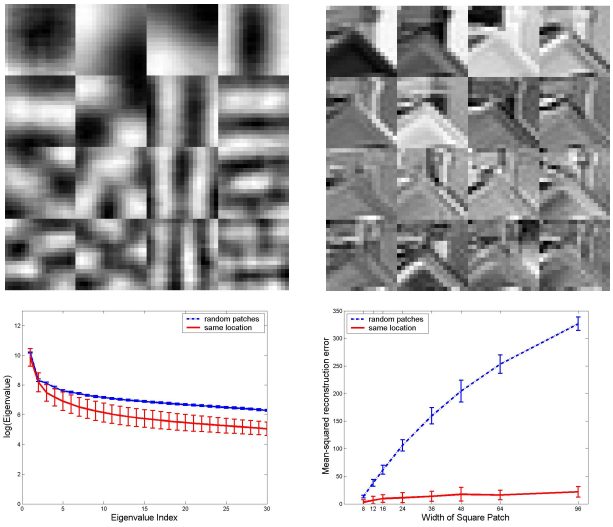


Figure 2. The covariant structure of multiple patches from the same location in a scene is much more informative than the covariant structure of arbitrary patches in natural scenes. This is apparent from the structure of the first 16 principal components, comparing (top left) patches taken at random locations from an image sequence, and (top right) patches taken from the same location in 1999 images from a static camera. This effect holds over many components and at many scales, the singular values of the SVD are significantly lower (in the log-plot on the bottom left), and the mean-squared error of the pixel intensity reconstruction stays nearly constant as the patch size increases.

In addition to a large amount of image data, each camera is assigned latitude and longitude coordinates; in most cases the coordinates are assigned by a human but in some cases the coordinates were estimated based on the camera IP address. The majority of cameras are located in the continental United States.

3. Natural vs. Location-specific Statistics

We compare the second-order statistics of natural image patches and location-specific patches. One goal of this is to discover how much benefit there is to make a representational basis that is specialized to a particular location, and to measure how this benefit scales with patch size. For all basis computations in this work we use the singular value decomposition (SVD) if memory permits and otherwise use incremental SVD [1].

In this section, we characterize the singular values of the SVD and reconstruction error for varying patch sizes and linear basis functions. For each camera in the AMOS database, we have approximately 2000 images taken during October and November. Location-specific statistics are created by randomly choosing a patch location, and for each image collecting the pixel values of that patch in a column vector I_j. We collect these vectors in a matrix

$\mathbf{I} = I_1, \ldots, I_n$ and compute the SVD $\mathbf{I} = U\Sigma V^\mathsf{T}$. We compute the reconstruction error for 200 randomly selected patch locations to determine the mean and standard deviation. The natural (non-location specific) image statistics were computed by selecting one patch from a random location (uniformly across the image) in each image of the scene. This naturally enforces the goal of using the same number of patches, and sampling from images throughout the day. This was also repeated 200 times, to determine the mean and standard deviations of the singular values and reconstruction error.

The results are shown in Figure 2; at the top are example results from (left) natural image patches, and (right) location-specific patches. The principal components of one set of 2000 natural image patches resemble a 2D frequency decomposition, as has been widely reported (see, for example [9]). These components look qualitatively similar between different repetitions. In contrast, the principal components for the location specific patches are drastically different from the natural components and between repetitions, because they reflect the structure of the scene in view at that location.

For a fixed patch size, the difference in the magnitude of singular values remains large out to as many values as we have computed, Figure 2 (bottom left) shows the mean and variance of the singular values for a 16×16 patch. The differences become even more dramatic for larger patch sizes. Using a fixed number of components (30), Figure 2 (bottom right) reports the mean and standard deviation of the mean-squared reconstruction error. This reconstruction error grows very slowly as a function of patch size, because most variations in appearance from a fixed camera are lighting changes that affect large parts the scene.

Since the reconstruction error remains small for large patches, we continue our analysis by considering principal components over the whole image. Here we take a 2000 frame sequence and compute both the principal components and the coefficients used to linearly reconstruct each image. Figure 3 shows for three example cameras, a sample image, the first three principal components, and the coefficient values plotted as a function of their time of day (color coded by which day). These coefficients are strongly correlated with time of day and are surprisingly similar between cameras. This highlights the fact that not only are the second-order statistics of static cameras interesting over large spatial scales, but there is also structure through time that is similar between multiple cameras. The remainder of this paper explores structure in the SVD of images at daily, and longer, timescales.

4. Daily Variations of Outdoor Scenes

In this section we explore temporal variations due to the time of day. To isolate variations due to transient phenomena such as weather and moving objects we construct an average day, a set of 48 average images, one for each half-hour of the day, from all of the images from the month of June 2006.

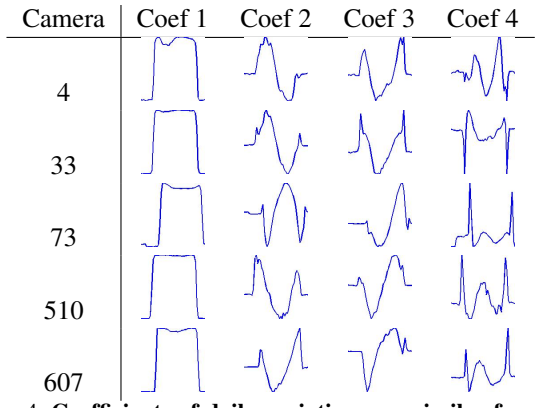


Figure 4. **Coefficients of daily variations are similar for static images of outdoor scenes.** Each row shows a plot of the first through fourth principal component coefficients for a different camera for an average day (a set of 48 half-hour average images from June 2006). The horizontal axis of each plot is the time of day and the vertical axis is the coefficient value. The coefficients for different cameras are similar despite the fact that the corresponding scenes are very different. The primary differences between the coefficients of different cameras are due to three factors: a shift due to local dawn/dusk time, permutation due to the relative strength of different types of variation in the scene, and inversion due to the SVD decomposition.

The SVD of this set of images highlights that while the the principal components are strongly dependent on the scene, the coefficient matrices V of different cameras are surprisingly similar. Figure 4 shows the first four principal component coefficients of several cameras, plotting columns of the coefficient matrix V_i which, by construction of our image set, corresponds to time of day.

The variations are due to three factors: a shift due to local dawn/dusk time, a column permutation due to the relative strength of different types of variation in the scene, and inversion due to the non-uniqueness of the SVD.

We temporally align the coefficient matrices V_i by considering the first column as a function of time and computing the extrema of its derivative (this corresponds to finding the rise and fall of coefficient one in Figure 4). All coefficient matrices are then linearly interpolated to have the same number of coefficients before dawn, during the day, and after dusk.

For cameras with known latitude and longitude, the standard deviation of our estimates when compared to standard civil twilight on June 15, 2006 was 19 minutes. This is a reasonable error value since there is only one image for every 30 minutes.

The remaining variation is in the order and sign of

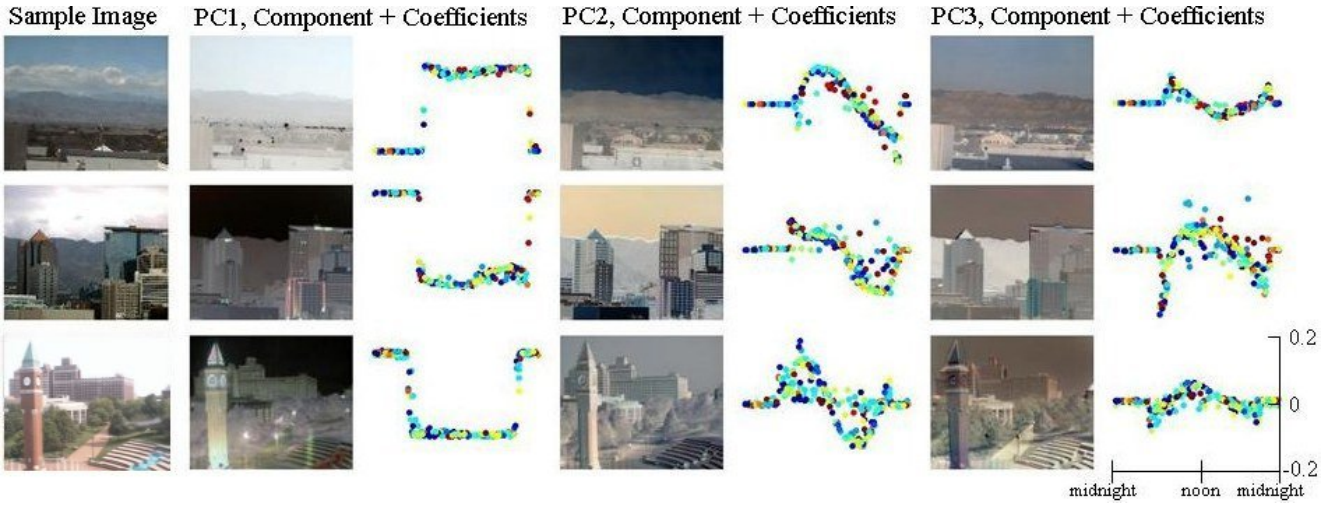
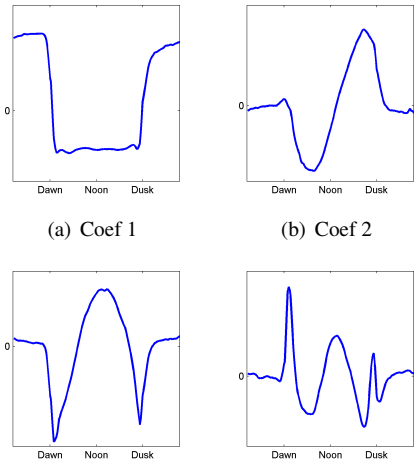


Figure 3. The most significant principal components of outdoor video captured by a static camera are often dependent on the time of day.



(a) Coef 1 (b) Coef 2

(c) Coef 3 (d) Coef 4

Figure 5. The first four canonical component coefficients learned from the AMOS dataset.

the columns of the coefficient matrices. Starting from temporally-aligned coefficient matrices V_i, we solve for a coefficient matrix $\bar{V}$ that is a solution to the following problem

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^{n} \max_{p \in \mathbf{P}} \left\| \bar{V}^{\mathrm{T}} p V_i \right\|_{\mathrm{F}} \\ \text{subject to} & \bar{V}^{\mathsf{T}} \bar{V} = \mathbf{I} \end{array}$$

where $\mathbf{P}$ is the set of generalized permutation matrices with entry values in the set $\{0, 1, -1\}$ and only one non-zero entry in each row and column. Figure 5 shows the first four canonical coefficients learned from a randomly selected set of 145 coefficient matrices.

Using $\bar{V}$, we can now decompose images from any camera, in a way that facilitates image and scene understanding. Given images $\mathbf{I_j}$ from camera j we can solve linearly for an orthogonal matrix of canonical components $\bar{U}_j$ and a diagonal matrix of weights $\bar{\Sigma}_j$ that are the solution to

$$\mathbf{I_j} = \bar{U}_j \bar{\Sigma}_j \bar{V}_j^{\mathsf{T}} \quad (1)$$

where $\bar{V}_j$ is $\bar{V}$ temporally-aligned to this camera.

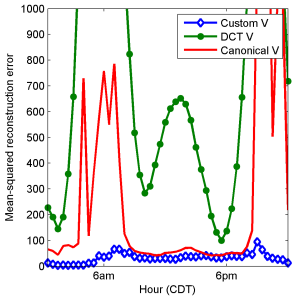


Figure 6. Average reconstruction error of half-hour average images for all cameras using three components.

The canonical day decomposition is not as good, in the squared error sense, at reconstructing images as the camera-specific SVD for the same number of components, but it is better than a generic low frequency decomposition (using DCT coefficients in place of V) by a factor of two. Figure 6 shows the reconstruction error by time of day.

The remainder of this section explores ways of directly comparing entries in the canonical day decomposition to aid scene and image understanding.

4.1. Image Labeling

Using the canonical day decomposition we can label individual images from the scene. Given any image I_j taken at time t from the scene we project it onto the canonical day components to obtain a vector of weights $c_j = \bar{U}_i^{\mathsf{T}} I_j$. These weights are then compared to the corresponding values, based on time of day, of the canonical day coefficients.

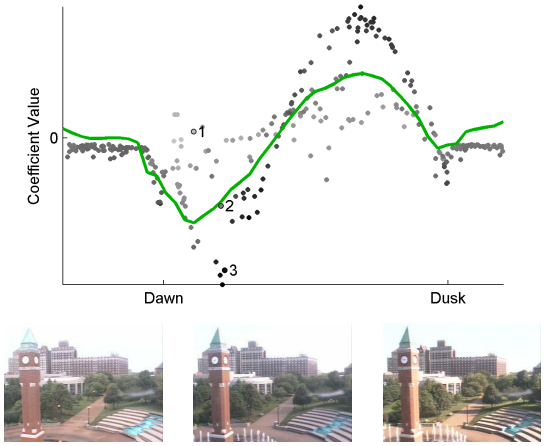
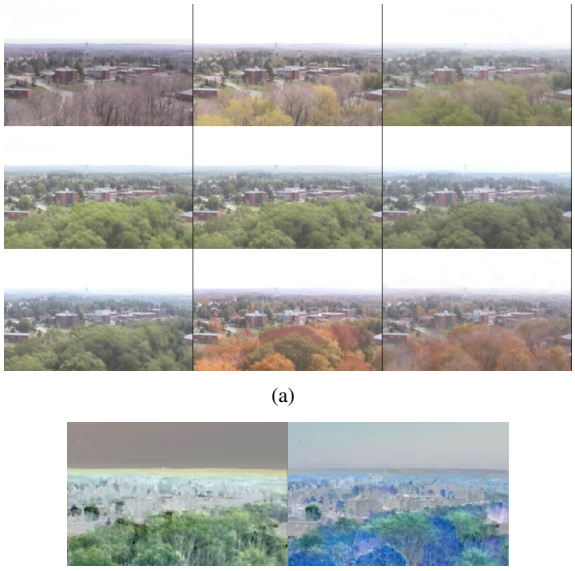


Figure 7. **Using canonical day decomposition to determine weather conditions.** The plot shows two types of coefficients for a set of images from a single camera. The solid line represents the values of the second canonical day coefficient (automatically aligned with dawn and dusk). The dots represent individual images from the camera with the x-value corresponding to the time when the image was captured and the y-value equal to the length of the projection of the image onto the second canonical day component. The dots are colored based on a function described in Section 4.1.



(a)

(b)

Figure 9. **Seasonal variations in outdoor scenes.** (a) A subset of 15-day-interval average images for an outdoor scene. (b) The first principal component (left) encodes primarily for the presence of trees, the second (right) for different types of trees.

As an example, Figure 7 shows a scatter plot of images colored by $c = |\bar{V}_i(t,2)\bar{\Sigma}_i(2,2)| - |\bar{U}_i(:,2)^{\mathsf{T}} I_j|$. This measure correlates with the cloudiness of the current image.

4.2. Scene Segmentation

One reason to create the canonical coefficient matrix is so that the canonical components that are computed following Equation (1) have a consistent meaning across all cameras. This allows one annotation scheme to be applied to all cameras. We create a false color image whose three color channels [R,G,B] are the third, second, and the negative of the first canonical components (this order was chosen so that strongly negative parts of the first canonical component are blue, mimicking the sky). This false color image strongly correlates with scene structure, and example images are shown in Figure 1, separating trees and horizontal surfaces (light green), eastward facing walls (orange/red), westward facing wall (blue).

5. Variations at Longer Timescales

We have shown that consistent patterns of daily variation occur across many cameras viewing a broad range of scenes. In this section, we show preliminary work exploring longer time scales. We find significant variations due to weather conditions, human activity, and the change of seasons.

We begin by examining variations that occur from day to day. In order to reduce effects due to the time of day we only look at images from one hour of the day. We create a set of 30 images for each camera, one for each day of June 2006, by averaging all images captured between 12:00pm and 1:00pm on each day. We then decompose this set of images using the SVD. Empirically, there is less regular temporal structure in this basis, in the V matrix, than in the basis of half-hour average images but the first component often has interesting structure. While the first component is often weather dependent it is occasionally caused by human activity (see Figure 8 for examples).

We now turn to variations at much longer time scales, here we focus on scene variations that occur at the scale of many months. Variations of this type include changes in shadow positions, changes in weather conditions (*i.e.,* snow on the ground), and changes in plants (*i.e.,* the presence or absence of leaves on deciduous trees). To reduce the effect of short term causes, we create a set of average images for each camera that includes primarily long-term variations. To do this we divide the year into 15-day intervals and create an image for each interval. The image is the average of all images within the interval captured between 12:00pm and 1:00pm. While we do not have sufficient data to draw meaningful conclusions, the SVD of this set of images is often highly structured (see Figure 9 for an example). As the AMOS dataset grows, we plan on developing decompositions that are similar to the canonical day decomposition but which are defined over much longer timescales.

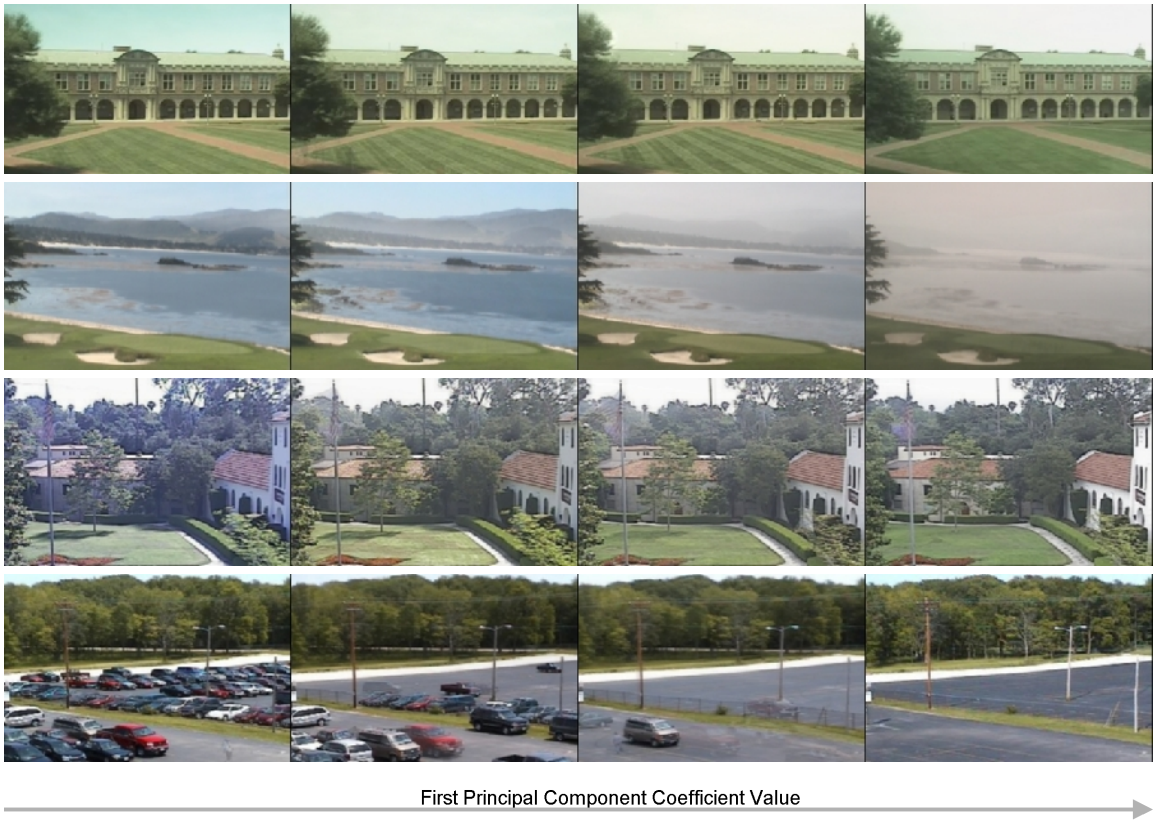


Figure 8. Images of day-to-day variations, described in Section 5, organized by the value of the first principal component coefficient. This value changes from day-to-day and is often dependent on the weather; occasionally it is dependent on human causes. The first principal component of the scene on the bottom is dependent on the presence of cars in the parking lot.

6. Conclusion

We have found that image sets from static cameras have strong correlations over large spatial and temporal extents. The principle components of these data sets, and their temporally-aligned variants, are useful because they can be compared between cameras and provide simple and automated tools to extract scene structure. We believe that understanding long scale spatial and temporal correlations in static video sequences is vital better understanding classic studies of the statistics of natural imagery. It may also directly affect the compression and transmission of surveillance video creation and maintenance of surveillance background models.

References

[1] M. Brand. Incremental singular value decomposition of uncertain data with missing values. In *Proc. European Conference on Computer Vision*, pages 707–720, 2002.

[2] D. W. Dong and J. J. Atick. Statistics of natural time-varying images. *Network: Computation in Neural Systems*, pages 345–358, 1995.

[3] P. Hancock, R. Bradley, and L. Smith. The principal components of natural images. *Network*, 3:61–70, 1992.

[4] S. J. Koppal and S. G. Narasimhan. Clustering appearance for scene analysis. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 1323–1330, 2006.

[5] S. G. Narasimhan and S. K. Nayar. Shedding light on the weather. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, volume 01, page 665, 2003.

[6] S. G. Narasimhan, C. Wang, and S. K. Nayar. All the images of an outdoor scene. In *Proc. European Conference on Computer Vision*, pages 148–162, 2002.

[7] B. A. Olshausen. Learning sparse, overcomplete representations of time-varying natural images. In *ICIP (1)*, pages 41–44, 2003.

[8] B. A. Olshausen and D. J. Field. Natural image statistics and efficient coding. *Network*, 7(2):333–340, 1996.

[9] E. P. Simoncelli and B. A. Olshausen. Natural image statistics and neural representation. *Annual Review of Neuroscience*, 24:1193–1216, 2001.

[10] 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*, 265:2315–2320, 1998.