COMPARING HIGHER-ORDER SPATIAL STATISTICS AND PERCEPTUAL JUDGEMENTS IN THE STYLOMETRIC ANALYSIS OF ART

James M. Hughes, Daniel J. Graham*+, C. Robert Jacobsen*†#, Daniel N. Rockmore*†**

*Department of Computer Science
† Department of Mathematics
+Department of Psychological and Brain Sciences
Dartmouth College
03755 Hanover, New Hampshire, USA
phone: + 1 (603) 646-8743, fax: + 1 (603) 646-1312,
email: rockmore@cs.dartmouth.edu

#Department of Mathematical Sciences, Aalborg University
Frederik Bajers Vej 7G, DK-9220
Aalborg East, Denmark

ABSTRACT

Understanding the factors that underlie human perception of artistic style necessarily depends on measuring features beyond simple, low-order statistics. Indeed, much of our perception of style is rooted in qualities of lines and shading that cannot be described using first-order measures. In order to provide a richer, more accurate description of human perception of style, we must employ higher-order statistical methods. We demonstrate the applicability of two types of higher-order representations of images – and features derived from these – to the problem of similarity-based image search in large collections of art images. Our preliminary results indicate that a combination of perceptual information and statistical representations of art could prove extremely useful in navigating large art image databases in the context of similarity-based search.

1. INTRODUCTION

Increasingly, large collections of visual art are being digitized and made available through the World Wide Web (see e.g., www.googleartproject.com), both for enjoyment by the general public and for analysis by the scientific community. Of interest in such viewing experiences is the possibility of navigating these large image spaces according to concepts of visual similarity. This can be a challenging goal to accomplish, however, because of the complex statistical structure of visual art, which shares many basic statistical regularities with natural scenes. For example, spatial frequency amplitude spectra (second order statistics) in art images have roughly the same $1/f$-shaped falloff found in natural scenes [8, 9, 20].

Because low-level statistics are so widely shared across art types (even abstract works typically have very similar amplitude spectrum slopes to those found in representational works [9]), additional measures are needed to develop large-scale systems for predicting similarity among paintings and drawings in large collections of artworks. Such systems would be of potential use for the applications mentioned above, and we suggest that the principles guiding similarity judgment in artwork will go a long way towards improving content-based image retrieval systems more generally.

We argue that characterizing regularities in higher-order statistics is a powerful approach to the problem of organizing large art databases according to stylistic similarity (and, indeed, perceptual similarity). This suggestion is in line with past work showing that machine learning algorithms trained on higher-order spatial statistics are effective at performing fine-grained stylometric[1] distinctions [12]. It also aligns with neuroscientific evidence that sparse statistical regularities [7] shape neural coding strategies in the visual system [5, 6]. That is, tailoring a “dictionary” of stylistic features to the higher-order redundancies found in art is akin to an organization schema of the receptive fields of visual neurons according to higher-order statistical regularities in nature [17].

Put another way, we efficiently adapt our representation to the higher order statistics of each image or image class, rather than using a standard orthonormal representation. This approach stands in contrast to the “kitchen sink” approach employed by other researchers [21], wherein one or more sets of features are chosen in an ad hoc way (e.g., RGB distributions, wavelet coefficients, face detection, etc.) to represent or describe a given image. While the latter approach has made important progress related to the analysis of large art databases – succeeding, for example, in separating art of different eras (e.g., Gothic vs. Impressionist) – there may be more principled ways to address the problems of quantifying style and using this information for image search and organization. Our approach, which includes making use of representations that are optimized for each image, attempts to provide a solution to these problems.

In addition, it remains to be seen whether style itself is a quality that is defined primarily with respect to perceived similarity as judged by lay viewers, historical or geographical provenance, or scholarly opinion. Psychological research in this vein – comparing style judgments in computer models trained on higher-order redundancies, art experts, and lay viewers – is underway. At present, though, there are but a handful of quantitative studies of the factors that govern human style perception.

Here we describe experiments demonstrating the effectiveness of representations that capture the higher-order statistics in a large, diverse collection of artworks. We com-

[1] “Stylometry” is a general term used to describe the development of quantitative tools for the analysis and understanding of artistic style.

pare this approach to other representational methods such as Gabor functions, as well as to approaches involving two-point statistics. We find that these statistical measures produce clusterings that agree in large part with a “true” underlying stylistic similarity between works of art, namely a labeling of the works based on the artist who created them. Further, we apply measures of higher-order redundancies to three sets of paintings of varying content (abstract art, landscapes, and portraits). A comparison of the resulting stylometric spaces to human judgements of similarity for the same image sets shows that approaches to stylometry using higher-order spatial regularities offer a promising route to capturing similarities across large, diverse collections of artworks.

We provide strong initial evidence that stylometric measures using higher-order statistical regularities show correspondences with perceived similarity. We believe that this strongly suggests that systems that base judgements on human perceptual information, while at the same time taking advantage of as much quantitative information as possible, are likely to provide the best performance in navigating image spaces using similarity-based search techniques.

2. IMAGES

The images we used were a collection of 308 high-resolution art images obtain from various sources. Included are drawings by Bruegel [19], paintings by Charlotte Caspers [3], paintings by Georges Braque [16], drawings by Raymond Pettibon [16], and a large collection of works spanning several centuries obtained through the Cornell University collections [8, 9], among others. All images were uncompressed TIFF or PNG images and were converted to grayscale via Matlab’s `rgb2gray` function before analysis [15].

3. IMAGE FEATURES

We examine the efficacy of two types of methods – fixed and adaptive – for providing descriptions of the stylistic qualities of art images. Furthermore, we compare these methods to the “expected” stylistic distinctions, as well as to psychophysical experiments that examined perceptual similarity between works of art [10]. The two image decomposition methods we utilize in this paper are a Gabor filter decomposition of images [4] and the sparse coding model [17, 18]. Several features are extracted from the decompositions obtained using each of these models and are described in more detail in the corresponding sections.

Gabor filter decomposition

Gabor functions are localized, oriented, and bandpass, and as such are sensitive to constructs of lines and edges at particular orientations and spatial frequencies. In our experiments, we created a set of Gabor functions at eight orientations (0 to $\frac{7\pi}{8}$ radians), four spatial frequencies (approximately $5, 9, 12$, and 16 cycles-per-picture), and two phases (0 and π radians), for a total of 64 filters.

Once an image patch size (e.g., 64×64 pixels) and filter size (e.g., 32×32 pixels) were determined, we imposed a grid on the images and extracted as many patches of the specified size as possible. Each of these patches was convolved with the Gabor filters we created to generate a set of 64 filter responses. Generally, we let the filters have a side length equal to one-half the side length of the image patches. This allowed us to obtain a section of the convolution image equal in size to the filter, disregarding parts of the image where zero-padding would have been necessary.

Once the response images for each patch were obtained, a feature vector was generated for each patch using the energy contained in each filter response:

$$E(I, f_{k,\theta,\phi}) = \sum_i |(f_{k,\theta,\phi} * I)[i]|^2,$$

where I is the image patch and $f_{k,\theta,\phi}$ is a Gabor filter with preferred spatial frequency k, preferred orientation θ and spatial phase ϕ, and i indexes pixels in the image patch. Other features are of course possible, but for our purposes here we considered only this method of feature extraction. Distances between works of art were determined by the correlation distance (i.e., $1 -$ Pearson’s r) between the average of the feature vectors associated with a particular image.

Sparse coding model

The sparse coding model of Olshausen & Field [17, 18], which is equivalent to independent component analysis (ICA) [1], was originally proposed to explain the response properties of cortical “simple cells” in the early visual system. The model learns a set of basis functions tuned to the higher-order statistical characteristics of a particular image space via maximum likelihood estimation. Since a sparse prior is used on the coefficients for any particular representation, the model attempts to maximize sparseness while guaranteeing a suitable level of reconstruction (i.e., one with relatively low reconstruction error).

For our purposes, we seek to take advantage of two important characteristics of this model: its sparseness, determined by non-Gaussian filter response distributions which allow the learned functions to be non-orthogonal and (possibly) overcomplete, and its adaptiveness, which insures that the learned functions are optimal with respect to the data. Sparseness is critical so that the functions do not become those that would be determined by a principal component analysis [2] decomposition of the image space, since such functions, which resemble the Fourier basis in two dimensions [17] and thus contain no localized information, are usually tuned to a narrow range of spatial frequencies and are generally not separable in terms of orientation and spatial frequency. Adaptiveness is also key: in contrast to a fixed decomposition such as a set of Gabor functions, the functions learned by the sparse coding model are data-dependent, and the variations in the properties of the functions themselves should be reflective of the underlying inputs.

Because of their adaptiveness to the input image space, we use the functions themselves as a proxy through which to analyze properties of the higher-order statistical characteristics of the images. Olshausen & Field showed that the learned functions reflect properties of the input image space [17]. We derive several features from the functions in order to analyze and compare these properties. In all of our experiments, we trained a set of 256 16×16 pixel basis functions on each image *individually* using the sparse coding model. It was from this set of functions that we derived features representing each image.

We compared images according to several metrics, which depend on the features extracted from the basis functions. They are as follows:

- Peak orientation: given the two-dimensional Fourier transform of a basis function, at what orientation does peak amplitude (or power) occur, averaged across all spatial frequencies. This is a reliable way of determining the orientation selectivity of a basis function.
- Peak spatial frequency: given the two-dimensional Fourier transform of a basis function, at what spatial frequency does peak amplitude (or power) occur, averaged across all orientations. This is a reliable way of determining the spatial frequency selectivity of a basis function.
- Orientation bandwidth: given the two-dimensional Fourier transform of a basis function, what is the bandwidth in octaves (measured by full width at half-maximum) of the function, averaged across all spatial frequencies, centered around its peak orientation. This quantity measures how selective a basis function is for its preferred orientation.
- Spatial frequency bandwidth: given the two-dimensional Fourier transform of a basis function, what is the bandwidth in octaves (measured by full width at half-maximum) of the function, averaged across all orientations, centered around its peak spatial frequency. This quantity measures how selective a basis function is for its preferred spatial frequency.

These quantities are computed for each of the 256 basis functions trained for each image. Since there is no natural way to compare individual functions with one another, we employ distributional methods to do so. In particular, we use symmetrized KL divergence (KLD) to compare distributions of these quantities, defined in the following way:

$$\mathrm{KLD}(P,Q) = \frac{1}{2}\sum_{\omega\in\Omega}\left[P(\omega)\log\frac{P(\omega)}{Q(\omega)} + Q(\omega)\log\frac{Q(\omega)}{P(\omega)}\right].$$

Since the values above are continuous quantities, we estimate KLD by binning, and we determine bins *once* for a particular quantity (e.g., spatial frequency bandwidth), and this determines the binning for all subsequent computations of the KLD. Thus, given the quantities above, we derive distances between all images using KLD for the following distributions:

- Distribution of peak orientation
- Distribution of peak spatial frequency
- Joint distribution of peak orientation and spatial frequency
- Distribution of orientation bandwidth
- Distribution of spatial frequency bandwidth
- Joint distribution of orientation and spatial frequency bandwidth

Furthermore, we compute distances between images based on a distance metric defined directly on the sets of basis functions [13]. The final feature we compute, from which we derive a distance, is the slope of the log rotational average of the amplitude spectrum for each image [5]. Ultimately, including Gabor filter energy, all of the distances derived from the sparse coding model basis functions, and the slope of the log rotational average of the amplitude spectrum, we have in total ten distances with which we compare images in our dataset. We include an eleventh, the distance matrix derived by aggregating the distance matrices after rescaling each so that the maximum distance was 1.

4. RESULTS

Our ultimate goal in this work is to provide a method for quantitatively characterizing the style of a work of art, specifically in such a way that allows us to easily compare a given work with others and determine which are stylistically most similar. Unfortunately, we do not at present possess a “ground truth” notion of stylistic similarity for our entire dataset. In order to compare the various derived metrics on images, we chose to compare them according to the true artist labeling of the images. For example, all paintings by Picasso would have the same label. This rule was used consistently except in one prominent case, the “non-Bruegel” category of drawings, in which drawings were given the same label, though they may be by different artists. This oversmoothing was necessary since the attribution of many of the Bruegel imitation images is not known; however, they are, like the Bruegel drawings, fairly stylistically consistent, especially with respect to the other works of art in the dataset.

In order to compare the information contained in the ten distance matrices with the true artist labeling of the images, we first embedded the drawings in Euclidean space via classical multidimensional scaling [11]. We then used the k-means algorithm to determine a clustering of the points for several values of k (shown in Figure 1). Once a clustering was determined for the embedding of each distance matrix, we compared that clustering with the true labeling using *normalized mutual information* (NMI) [14]:

$$\mathrm{NMI}(\Omega,C) = \frac{I(\Omega,C)}{[H(\Omega)+H(C)]/2},$$

where I is the mutual information between clustering Ω and the true labeling C,

$$I(\Omega,C) = \sum_k\sum_j P(\omega_k\cap c_j)\log\frac{P(\omega_k\cap c_j)}{P(\omega_k)P(c_j)},$$

and H is the entropy of each set of objects,

$$H(\Omega) = -\sum_k P(\omega_k)\log P(\omega_k).$$

Dividing by the average of the entropies ensures that this quantity is between 0 and 1. Although this measure is not a perfect means of quantifying the accuracy of a clustering (which is difficult in a case where labels are not identifiable), it does provide a reasonable means of estimating the information overlap of two clusterings, and we believe that this is an acceptable proxy for our purposes in these initial experiments.

Figure 1 shows the value of NMI for clusterings obtained via k-means clustering using each of the distance metrics described above, for several values of k. As can be seen, the best performance was obtained using the combined (i.e., aggregated) distance matrix, with the Euclidean distance-based basis metric, Gabor filter energies, and joint orientation/spatial frequency bandwidth distributions also containing information that was consistent with the true labeling of the images. The overall success of the combined distance matrix suggests that incorporating statistical measures based on fixed representations (such as Gabor filter energies and the slope of the amplitude spectrum) as well as adaptive measures (like those that depend on basis functions learned via a sparse coding model) is the most effective way to characterize stylistic properties in works of art.

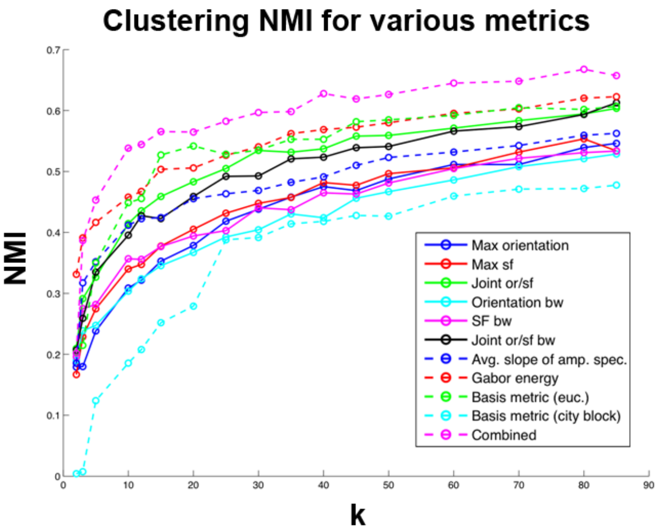

Figure 1: Clustering normalized mutual information for the 308 images used in our experiments, clustered using the k-means algorithm, across several values of k (indicated by open circles on traces in plot). A maximum of 85 clusters was chosen, since that corresponds to the number of unique artist labels in the dataset. A perfect clustering would have NMI of 1.

Comparison to perceptual experiments

In order to begin to evaluate the effectiveness of the derived features at accounting for *perceptual* similarities between images, we compared our feature-based distances to those obtained through psychophysical perceptual similarity experiments [10]. In these experiments, participants were asked to judge the similarity between pairs of art images (1-9 scale) in three categories, abstract art, landscapes, and portraits (images were sorted in a prior three-alternative forced choice test by a separate set of subjects). This information was aggregated across all participants to create a similarity matrix for each category. These experiments were extremely small scale and dealt with approximately 20 images per category, nevertheless, as we will demonstrate, the perceptual similarities between works of art provide information that is effective at categorizing images according to their style, at least at the coarse artist-by-artist scale of our experiments.

The first experiment we performed was to compare the effectiveness of the perceptual judgements at predicting the stylistic relationship between works of art. We accomplished this by holding out two images from each of the three sets of images. We then trained a regression model on the perceptual distances using the feature-based distances as regressors, except for the aggregated distance matrix (in our model, we include constant, linear, and quadratic terms). Using regression in this manner, we should be able to predict the relationship between two images, according to the perceived similarity between them, assuming the perceptual model contains useful information that will be predictive for the held-out images. We computed predictor weights in a regression model 500 times for each image category, holding out two different images at random each time. After each model was learned, we predicted the distances between the held-out images and all images used in training. We then selected one of these at random and compared the true perceptual distance relationship between the images to the relationship between the predicted distances. Accuracy and statistical significance of these results is shown in Table 1.

Category	Accuracy	p-value
Abstract art	0.61	5×10^{-7}
Landscapes	0.62	1×10^{-8}
Portraits	0.51	0.28

Table 1: Accuracy of model at predicting relationship of held-out images to randomly selected image in training set. High accuracy implies that the correct relationship between the test images (e.g., image 1 was closer to target image T than image 2 was) was predicted by the model. Accuracy is given by the fraction of correct predictions (out of 500 tests). The right column shows p-values indicating the significance level of these tests, assuming a binomial model in which the correct relationship would be guessed at random (i.e., by flipping a fair coin).

These results indicate that perceptual information from two of the three categories (abstract art and landscapes) contained information that allowed prediction at a statistically significant level. This not only confirms the existence of useful information in the perceptual similarity data, but also the ability of the statistical information to effectively model these distinctions. Nevertheless, the number of images used in these experiments was extremely limited, and in order to generalize these results, further experiments are required.

We also explored the extent to which the admittedly limited perceptual information we possessed about the three categories of images could be predictive of stylistic distinctions in larger sets of images. This application is of particular importance to similarity-based image search, since its success implies that limited subsets of perceptual similarity information between images could be used to “bootstrap” models used to predict similarity between very large sets of images.

Prediction of perceptual distances between images in our dataset was accomplished by first modeling the perceptual distances in the three image categories using the feature-based scores, as before, but this time without holding out any images. Once a model was obtained, we averaged the the predictor coefficients from each of the three models and used this to predict the “perceptual distance” between images. Note that, in these experiments, we did not predict distances for the images that were included in the original perceptual experiments (i.e., those in the abstract, landscape, and portrait categories). These images were held-out during the prediction phase and did not factor into the subsequent analysis.

We held out 51 images (across the three categories) and created a predicted perceptual distance matrix on the remaining 257 images. We compared this with the same distance measures used above by using only the submatrices that contained these 257 works in each of the 11 original matrices. As before, we found a Euclidean embedding of the images using each distance matrix and then performed k-means clustering for several values of k, then compared these clusterings with the corresponding true labeling. The results are shown in Figure 2. Not surprisingly, the aggregate distance matrix still yields clusterings with the closest relationship to the true

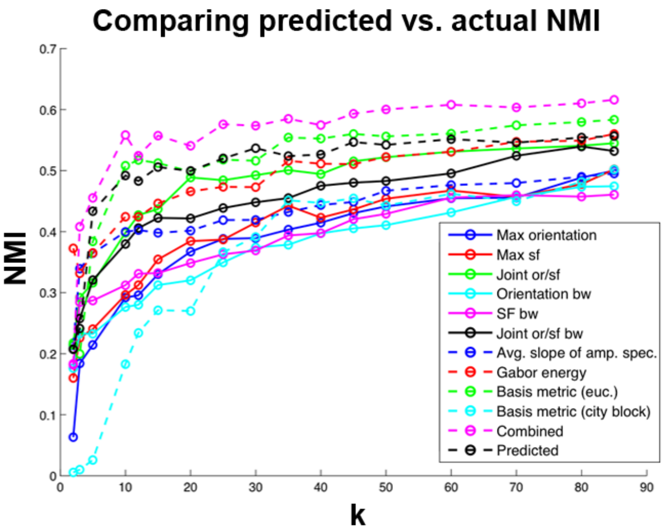

Figure 2: Clustering normalized mutual information for 257 images used to compare clusterings between distance measures and predicted perceptual distances, clustered using the *k*-means algorithm, across several values of *k* (indicated by open circles on traces in plot). A perfect clustering would have NMI of 1.

labeling. However, the predicted distance matrix has performance on par with any other individual feature, suggesting that, although our sample of perceptual distances was extremely limited, this information can guide the ways in which we combine statistical features to understand style perception.

5. CONCLUSIONS

Although preliminary, our results indicate not only that measurement of higher-order statistical characteristics of images generates information germane to stylistic distinctions, but also that combining this information with perceptual similarity can create an effective, statistical means of organizing images and predicting perceptual similarity in the context of similarity-based search in large image databases. Future work will include implementation of such a system and further analysis of the concepts presented here.

6. ACKNOWLEDGEMENTS

JMH, DJG, and DNR gratefully acknowledge the partial support of the Samuel H. Kress Foundation for this work.

REFERENCES

[1] A. Bell and T. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, no. 6, pp. 1129–1159, 1995.

[2] C. M. Bishop (2007) *Pattern Recognition and Machine Learning*, New York: Springer, 2007.

[3] Charlotte data set. http://www.math.princeton.edu/ipai/datasets.html.

[4] J. Daugman, "Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters," *Journal of the Optical Society of America A*, vol. 2, no. 7, pp. 1160–1169, 1985.

[5] D. J. Field, "Relations between the statistics of natural Images and the response profiles of cortical cells," *Journal of the Optical Society of America A*, vol. 4, pp. 2379–2394, 1987.

[6] D. J. Field, "What is the goal of sensory coding?" *Neural Computation*, vol. 6, pp. 559–601, 1994.

[7] D. J. Graham and D. J. Field, "Sparse coding in the neocortex," In *Evolution of Nervous Systems* , Vol. III, eds. J. H. Kaas and L. A. Krubitzer. Oxford: Elsevier, pp. 181–187, 2006.

[8] D. J. Graham and D. J. Field, "Statistical regularities of art images and natural scenes: spectra, sparseness and nonlinearities," *Spatial Vision*, vol. 21, pp. 149–164, 2007.

[9] D. J. Graham and D. J. Field, "Variations in intensity statistics for representational and abstract art, and for art from the eastern and western hemispheres," *Perception*, vol. 37, pp. 1341–1352, 2008.

[10] D. J. Graham, J. D. Friedenberg, D. N. Rockmore and D. J. Field, "Mapping the similarity space of paintings: image statistics and visual perception," *Visual Cognition*, vol. 18, pp. 559–573, 2010.

[11] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, New York: Springer, 2005.

[12] J. M. Hughes, D. J. Graham, and D. N. Rockmore, "Quantification of artistic style through sparse coding analysis in the drawings of Pieter Bruegel the Elder," *Proceedings of the National Academy of Sciences*, vol. 107, pp. 1279–1283, 2010.

[13] C.R. Jacobsen, J.M. Hughes, D.J. Graham, and D.N. Rockmore, "A distance metric between sets of vectors," *in preparation*.

[14] C. Manning, P. Raghavan, and H. Schütze, *An Introduction to Information Retrieval*, Cambridge: Cambridge University Press, 2009.

[15] Matlab software. The Mathworks, Natick, MA, 2011.

[16] Museum of Modern Art, New York City. Courtesy Jim Coddington.

[17] B. Olshausen and D. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, pp. 607–609, June 1996.

[18] B. Olshausen and D. Field, "Sparse coding with an overcomplete basis set: a strategy employed by V1?," *Vision Research*, vol. 37, no. 23, pp. 3311–3325, 1997.

[19] Orenstein, N. M., Ed. *Pieter Bruegel - Drawings and Prints*. Metropolitan Museum of Art, New York City and Yale University Press, New Haven, 2001.

[20] C. Redies, J. Hasenstein and J. Denzler, "Fractal-like image statistics in visual art: similarity to natural scenes." *Spatial Vision*, vol. 21, pp. 137–148, 2007.

[21] C. Wallraven, R. Fleming, D. W. Cunningham, J. Rigau, M. Feixas and M. Sbert, "Categorizing art: comparing humans and computers," *Computers and Graphics*, vol. 33, pp. 484–495, 2009.