Naturalness and Memorability: A Study of Image Similarities and Memorability in Natural and Urban Images

Yutai Li

University of Chicago

Author Note

May 2023

A paper submitted in partial fulfillment of the requirements for the Master of Arts degree in the Master of Arts in Computational Social Science

Faculty Advisor: Marc Berman

Preceptor: Jon Clindaniel

Abstract

Interacting with nature is beneficial. Even brief exposures to natural environments or viewing sets of natural scene images can restore attention and memory (Berman et al., 2008). Despite the known benefits of nature, it remains unclear what aspects of viewing nature scenes lead to these improvements. While previous studies focus on the distinction between natural images and non-natural (urban) images, very few examine the connection between distinct properties of natural images and their cognitive benefits. One idea is that natural scenes are more fluently processed, making them less memorable, which in turn contributes to their benefits. A potential confound is that nature scenes may be more confusable than urban scenes, leading to their lower memorability scores predicted by the ResMem Neural Network (Needall & Bainbridge, 2021). The goal of this thesis was to test the hypothesis that image set similarity drives differences in memorability.

To test this hypothesis, the present study calculated image statistics and image semantic labels into vectors and determined the Euclidean and cosine distances between pairs of natural and urban images. While the low-level visual features in natural and urban images differ, importantly, there were no significant differences in the average distances of every image pair. However, the semantic similarity analysis showed mixed results. The significant differences in semantic similarity were shown in both direction under different types of label coding techniques. This finding suggests the relationship between image similarity and memorability is more complex than a simple positive or negative relationship. In the future, studies can focus on the processing fluency and memorability of natural images to uncover the mechanism behind their restorative effect.

Naturalness and Memorability: A Study of Image Similarities and Memorability in Natural and Urban Images

Introduction

The diminishing interactions with nature are depriving individuals of the myriad cognitive benefits that natural scenes have to offer. Since the mid-20th century, the number of people living in urbanized cities has doubled worldwide (United Nations, 2014). With convenient access to robust infrastructure and diverse entertainment, metropolitan residents are becoming increasingly bounded within cities and disconnected from nature. However, the benefits of connecting with nature are non-negligible. Numerous studies have shown that neighborhoods with high exposure to green space are associated with better health outcomes, such as fewer cardio-metabolic conditions (Karden et al., 2015), and reduced criminal activities (Schertz et al., 2015). Individuals connected to nature also experience various cognitive benefits. Even brief exposure to the natural environment or viewing pictures of natural scenes has beneficial effects on attention, memory, and mood (Berman et al., 2008; Berman et al., 2012). With all the benefits of nature, the distinct properties of natural environments have become a topic that attracts significant attention from researchers.

The beneficial effects of nature are explained by three dominant theories: biophilia, stress reduction, and attention restoration (Capaldi et al., 2015). The Biophilia hypothesis refers to human beings' evolutionary preferences toward nature (Kellert & Wilson, 1993). The stress reduction theory (Ulrich et al., 1991) emphasizes that unthreatening natural environments are evolutionarily preferable, which generate stress-reduction psychophysiological responses. The attention restoration theory (Kaplan & Kaplan, 1989), from the cognitive perspective,

distinguishes between two types of attention: directed attention and involuntary attention. In natural settings, directed attention is minimized because interacting with nature is not cognitively demanding. In urban settings, however, both involuntary and directed attention are significantly invoked to perform tasks such as avoiding traffic (directed) and hearing noises (involuntary). Compared to urban settings, natural environments enable directed attention to replenish and restore. In Berman et al.'s study (2008), participants who took a long walk in a park showed significant improvement in the performance of the backward digit span task, which requires directed attention and short-term memory. Participants who viewed a set of natural scene images also showed significant improvement in the Attention Network Test (ANT), demonstrating that interacting with nature and seeing natural images have restorative effects on attention (Berman et al., 2008).

What is intriguing about these findings is that if viewing natural images as an isolated visual input also leads to cognitive benefits, the properties of natural images might be inherently different from non-natural images, and these distinct properties may be related to their cognitive benefits. Since the last century, statistical modeling has been used to obtain mathematical representations of image patterns for analytical purposes. Through statistical modeling, one notable property of natural images is their invariance to scale, meaning that the low-level characteristics of the image remain relatively consistent when zooming in and out of a scene (Ruderman, 1994). This finding shows that natural images embody specific properties that non-natural or artificial images do not have, which implies some associations between their distinct characteristics and the cognitive benefits of viewing natural images. In addition, low-level image statistics contain more information than people previously expected. Torralba and Oliva's study found that natural scene categories (mountain, beach,

forest, etc.) and the presence of natural objects in a scene can be predicted directly based on some low-level visual features like spectra of natural images and scale of scenes (Torralba & Oliva, 2003). Their finding also supports the hypothesis that the low-level features between natural and non-natural images are inherently different.

However, one cannot make an arbitrary conclusion that the distinct statistical regularities of natural images are related to their beneficial cognitive effects for several reasons. First, the definition of natural images was not uniform in earlier research. Some studies defined them as "outdoor images of nature," while others provided a narrower definition, such as "wooded environments in springtime" (Ruderman, 1994; Srivastava et al., 2013). Second, the purposes of these image statistics studies were mostly to classify image scenes and recognize objects. Previous findings from image statistics studies only support the first half of the argument that the image visual properties are different between natural and non-natural images. The connection between image visual features and their cognitive effects has not been established.

To extend our understanding of natural images, recent studies have found that both low-level and high-level features can predict the perceived naturalness of images. In Berman et al.'s study (2014), researchers used spatial multidimensional scaling (MDS) as well as participants' subjective ratings to classify images based on similarity without prior assumptions. Both ratings show a consistent perception of what is considered to be natural images, which eliminates the inconsistent definition of natural images. Then, an image classifier was built based on ten low-level features of the images to predict the perceived naturalness. The findings suggest that first, naturalness can be quantified either implicitly using a multidimensional scaling analysis or explicitly with perceived naturalness ratings. The ambiguity in classification is minimal; second, four low-level visual features: density of contrast

changes, density of straight lines, average color saturation, and average hue diversity are significant predictors when determining whether an image is a natural or urban image (Berman et al., 2014), suggesting these four low-level features are significantly different between natural and urban images.

The perceived naturalness can also be predicted by high-level semantic features of images. To further extend the research on perceived naturalness, Hyam's study (2017) found that the perceived naturalness can be predicted by the Google Vision API, an automatic image label detection algorithm, suggesting that the auto-generated labels also reflect the semantic information that corresponds to the perceived naturalness of images. Moreover, Kotabe et al.'s study has shown that even when only presenting the low-level features of natural images (e.g., edges), the high-level semantic features like 'naturalness' and 'disorder' can be preserved. When only presenting the color visual feature, the naturalness rating of the low-level image significantly correlates with the naturalness rating of the original images (Kotabe et al., 2016). This finding challenges the traditional assumption that high-level semantics are perceived after integrating low-level visual features. It also supports the notion that the perceived naturalness of an image is embedded in both low-level visual features and high-level semantic features, making natural images inherently different from urban images in both dimensions.

In addition to perceived naturalness, the perceived similarities between images are embedded in the low-level features as well. In Neumann and Gegenfurtner's study (2006), researchers investigated three low-level feature representations: chromaticity histogram (the color distribution of an image), luminance histogram (average luminance level of pixels in each color), and Fourier index (representing orientation and high-frequency contrasts). The luminance histogram and Fourier index used

Euclidean distance norm to calculate the distance between two images. The distance between two chromaticity histograms was calculated by the sum of the minimum of the corresponding color bin frequency. To measure the judgment of perceived similarity, they used the two-alternative forced-choice (2AFC) design. In the experiment, three images were presented: the query image at the top with two test images below, and participants selected which test image was more similar to the query image. The probability of choosing an image given a query image is defined as the perceived similarity between these two images. The correlations between the distances in low-level feature representations and the human-judged similarities were used to determine the relationship between low-level visual features and perceived similarity. The results indicated that human judgments were best predicted with the chromaticity histogram. The luminance and Fourier histograms both contribute to the similarity judgments, and the percentage of agreement increases considerably if the luminance and Fourier information are combined with the chromaticity index (Neumann & Gegenfurtner, 2006).

While previous studies have provided a clear definition of natural images and the information embedded in their low-level and high-level features, the question of whether these features lead to beneficial cognitive effects remains unanswered. A recent study approaches this question by proposing that the distinct image properties evoke positive associations that lead to beneficial cognitive effects. Specifically, the study focuses on the role of low-level visual features in automatic responses to nature and urban images (Menzel & Reese, 2021). In the study, the Implicit Association Test (IAT) is applied to test the automatic responses to original and low-level nature and urban images in three dimensions: valence (good, bad), mood (positive, negative), and stress vs. restoration. The results indicate that original natural images

are associated with automatic positive responses across all three dimensions, and original urban images are associated with automatic negative responses. However, when presenting images with the spatial information deleted but certain image properties retained (i.e., phase-scrambled images), neither natural nor urban images are strongly associated with positive responses; no significant difference was found except in the mood dimension with a small effect size (Menzel & Reese, 2021). This finding indicates that images without spatial information seem to have little effect on associations with positive effects like mood and stress restoration. High-level visual processing of images appears necessary for its associated cognitive benefits.

However, Menzel and Reese's study does not resolve the question about image features and cognitive effects for mainly two reasons. First, the experiment uses phase-scrambled images that mainly preserve color information, while other visual features like edges are altered. It is worth noting that the density of edges is also a significant predictor of perceived naturalness, and the natural component in the image may be destroyed when altering the edges feature. Second, the Implicit Association Test is controversial in terms of its validity, and previous studies have shown that using IAT to test unconventional classifications might not be reliable overall (Greenwald et al., 2005). When the spatial information is removed, the classification of natural and urban images becomes an unconventional task within the scope of IAT. None of the target categories (phase-scrambled images in this case) have any pre-existing bias with the attribute because of the lack of spatial information, so the application of IAT is not ideal for validating the automatic responses in low-level feature images.

The present study aims to extend the literature on the connection between natural images' distinct properties and their cognitive benefits. Given the differential

memorability scores of natural and urban images predicted by the ResMem Neural Network (Rim, et al., 2023), the present study focuses on whether the image similarity is an influential factor for images' memorability. Specifically, the hypothesis of the present study proposes is that natural images share similar image properties (low-level) and semantic concepts (high-level), making them perceived more similar so as less memorable than urban images. To compare the perceived similarities, the present study transforms every image into vectors in both low-level and high-level dimensions. In the analysis of low-level visual features, images are transformed into twelve dimensional vectors based on corresponding visual features. In the analysis of high-level semantic features, semantic labels and corresponding confidence scores are generated by the Google Vision API label detection algorithm, which has also been used in Hyam's study (2017). The semantic labels are transformed into 300-dimensional vectors based on the pre-trained Word2Vec model. Then, the study calculates the Euclidean and cosine distances between each pairwise images and compares the distance matrix of natural images with that of urban images. A smaller average pairwise distance indicates the image set is more similar within the image category, and a larger average distance means the image set is less similar and more diverse.

The present study contributes to the previous literature of natural images and cognitive effects by proposing a possible confound that nature scenes are more similar as a set, which may cause them to be less remembered as suggested by previous study (Rim, et al., 2023). The difference in perceived similarity between natural and urban images is a variable that previous studies neglected. In addition, the present study provides clear directions for future research to extend the present literature in several aspects. Perceived similarity is closely related to processing fluency (Hiatt & Trafton,

2013), and processing fluency has been shown to exert strong influences on affective judgment and familiarity (Westerman et al., 2015). If natural images as a set is perceived similar, natural images are potentially processed more fluently, which in turn suggests the instinctive preferences toward natural images. The present study also narrows the gap between image properties and cognitive benefits in a way that if natural images are perceived similarly, they may require less cognitive resources to be perceived and are able to replenish the memory and attention, which is known as the restorative effect. Visual attention is a strong predictor of image memorability and visual attention is related to the low-level visual features of images (Mancas & Le Meur, 2013). If natural images require less attention, they may also be less memorable due to the insufficient allocation of visual attention.

Empirical Analysis 1: Low-level Visual Features

The first part of the analyses focuses on the perceived similarities of natural images and urban images in low-level visual features. Low-level visual features refer to the basic properties of an image that humans perceive without further processing. There are three main types of low-level visual features: color-based features, texture-based features, and shape-based features (Bo & Yi, 2014). In this section, the study uses the dataset from Kotabe et al.'s study (2016), which contains twelve low-level features as well as the perceived naturalness rating of every image. The whole image set is divided into two equal-sized subsets: natural image set and urban image set by naturalness ratings, and every image is transformed into a twelve-dimensional vector based on twelve low-level visual features. The distances between every pair of images within the category are calculated, and the distance matrices are compared between natural images and urban images.

Method

Materials

The present study acquired the primary dataset from Kotabe et al. (2016) study and aims to take a supplementary approach to extend the previous finding. The dataset is publicly available on the website of the Environmental Neuroscience Lab of the University of Chicago. The dataset includes 1030 natural and urban images from the Scene UNderstanding (SUN) image database (http://vision.princeton.edu/projects/2010/SUN/; Xiao et al., 2010) with a uniform size (1200 * 900) in .jpg format. Twelve low-level visual features of these 1030 images are included in the dataset as well. The low-level features include six color-based features: average hue density (Hue; the degree to which a stimulus can be described as similar to or different from stimuli that are described as red, green, or blue), standard deviation of hue (SDhue; average standard deviation of hue across all of an image's pixels), average saturation (Sat; the degree of dominance of hue mixed in the color), standard deviation of saturation (SDsat; average standard deviation of saturation across all of an image's pixels), average brightness (Lum; average brightness across all image pixels), standard deviation of brightness (SDbright; average brightness across all image pixels), and six spatial-based features: entropy (Entropy; the average "information content of the image), straight edge density (SED: number of pixels on straight edges), non-straight edge density (NSED; number of pixels on non-straight edges), density of edges (ED; sum of straight and non-straight edges), vertical reflectional asymmetry (LRSymm; how well the left and right halves of the scene image mirror each other), and horizontal reflectional symmetry (UDSymm: how well the up and down halves of the scene image mirror each other). The color-based features are calculated based on the hue (saturation, brightness) of every pixel in an image, and the average hue (saturation, brightness)

across all image pixels, and the standard deviation of hue (saturation, brightness) is calculated as color-based low-level features. Entropy is defined as the corresponding states of intensity level that individual pixels can adapt, which shows an image's average information content. The edges of images are extracted by the Canny edge detection algorithm (Klette & Zamperoni, 1996), and the gradient-based connected component algorithm is used to detect straight and non-straight edges (For detailed definitions of the low-level visual features, please refer to Berman et al., (2014) study, p. 11-15 and Kotabe et al., (2017) study).

The ranges and the variances of twelve low-level visual features are not uniform. Table 1 lists the distributions and the variances of the twelve low-level features. The vertical reflectional asymmetry (LRSymm) and horizontal reflectional asymmetry (UDSymm) have significantly smaller magnitudes, and Entropy has a different scale than the other variables. Due to the unequal variances among low-level features, the present study calculates the distance matrices on the original data as well as on data after normalization to achieve uniform variances. Table 2 lists the descriptive data of low-level features after scaling. Also, 6 of the 1030 images have missing values in some low-level features, and they are discarded when obtaining descriptive data as well as in later analyses.

	Hue	Sat	Lum	sdHue	sdSat	sdBright	Entropy	LRSymm	UDSymm	SED	NSED	ED
Var.	0.01	0.02	0.01	0.01	0.00	0.00	0.15	0.00	0.00	0.00	0.00	0.00
Min.	0.06	0.02	0.14	0.00	0.04	0.05	5.59	0.00	0.00	0.00	0.00	0.04
1st Qu.	0.26	0.20	0.48	0.17	0.14	0.21	7.24	0.01	0.00	0.03	0.05	0.10
Median	0.33	0.29	0.54	0.22	0.19	0.25	7.49	0.01	0.01	0.05	0.08	0.12
Mean	0.34	0.30	0.54	0.21	0.19	0.25	7.39	0.01	0.01	0.06	0.08	0.12
3rd Qu.	0.42	0.38	0.60	0.26	0.23	0.29	7.66	0.01	0.01	0.08	0.11	0.14
Max.	0.77	0.95	0.96	0.46	0.43	0.42	7.96	0.02	0.02	0.21	0.19	0.20

Table 1. Descriptive data of low-level features before normalization

	Hue	Sat	Lum	sdHue	sdSat	sdBright	Entropy	LRSymm	UDSymm	SED	NSED	ED
Var.	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Min.	-2.41	-2.10	-4.05	-2.83	-2.35	-3.56	-4.62	-2.23	-1.89	-1.79	-2.18	-2.89
1st Qu.	-0.67	-0.75	-0.58	-0.53	-0.76	-0.61	-0.40	-0.77	-0.73	-0.76	-0.78	-0.69
Median	-0.09	-0.13	0.01	0.08	0.03	0.08	0.25	-0.07	-0.18	-0.20	0.06	-0.05
Mean	0.00	-0.00	-0.00	0.00	-0.00	-0.00	0.00	0.00	0.00	0.00	-0.00	0.00
3rd Qu.	0.65	0.59	0.61	0.63	0.67	0.64	0.68	0.71	0.57	0.62	0.69	0.66
Max.	3.63	4.77	4.16	3.27	3.76	2.95	1.45	3.13	3.80	4.84	2.87	3.02

Table 2. Descriptive data of low-level features after normalization

The perceived naturalness ratings of images are human ratings collected from Berman et al.'s study (2014), experiment 3. The procedure for collecting the perceived naturalness ratings involved showing participants a single image at a time and having them rate it on a scale of 1 to 7 for how natural they considered the image to be. A '1' indicated that the participants considered the image to be very manmade, and '7' indicated that participants considered the image to be very natural. A '4' indicated that the image was not judged to be either natural or manmade (Berman et al., 2014). The average perceived naturalness rating of all images is 4.265, and the median rating is 4.750. The present study uses the median of perceived naturalness ratings instead of the mean to divide natural and urban images because having two equal-sized image sets is more desirable in the analyses. If the numbers of images in two image sets are different, the numbers of pairwise distances in each image set will be different, which could potentially introduce more biases to the analyses.

Procedure

All images, except for six with missing values, are transformed into twelve-dimensional vectors, representing twelve low-level visual features of images. Then, Principal Component Analysis (PCA) is applied to reduce the dimensionality of the image vectors and visualize them in a two-dimensional plot. The two most significant components that account for the most variances are the x and y axes of

the graph. While the principal component analysis plots the distribution, it does not quantify the similarity between images. After visualizing the distribution, the distances of every pair of images within the natural and urban image sets are calculated. Each natural and urban image set contains 512 images and 130,816 observations of Euclidean and cosine distances (512*511/2 = 130,816). The present study uses both Euclidean and cosine distance norms to calculate the similarity between images to avoid biases in distance norms. The distributions of both distance matrices are plotted to compare the perceived similarity between natural and urban images.

Results

As the principal component analysis suggests in Figure 1, the distributions of nature images (red dots) and urban images (green dots) are relatively separated. Urban images are centered in the lower left of the figure, whereas natural images are centered at the central-right side of the figure. The low-level feature that explains the most variance is non-straight edge density (NSED), which accounts for 31.6% of the total variation in the dataset. This result matches Berman et al.'s (2014) study, which found that non-straight edge density is the most significant predictor when determining whether an image is a natural or urban image. In addition, the distribution of urban images is visually more clustered, whereas the distribution of natural images is more scattered with outliers.

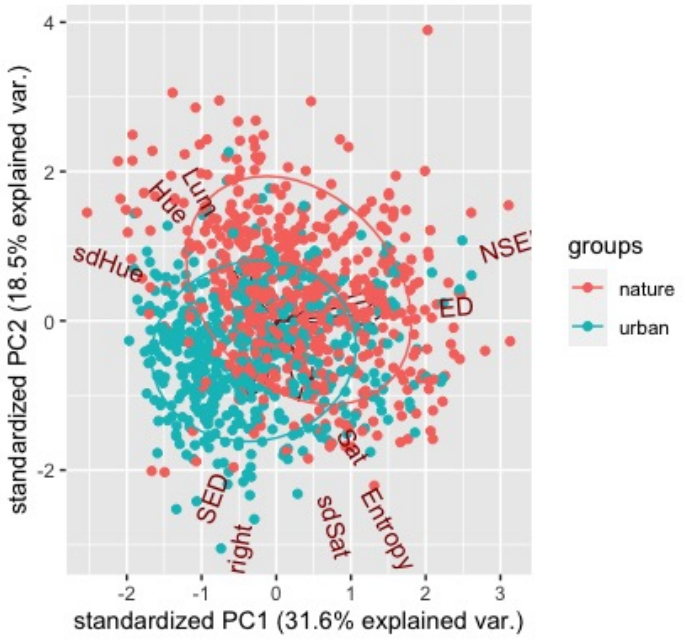

Figure 1. Principal Component Analysis (PCA) of low-level features

However, the principal component analysis does not quantify the distances in low-level visual features between natural images and urban images. Thus, the present study further calculates the Euclidean and cosine distances between every two images in the natural and urban image sets to determine whether natural images share similar low-level visual features. Figure 2a shows the Euclidean distance norms before using standardized low-level visual features. As Figure 2a indicates, both distributions of natural and urban images skew to the right, meaning that most images within the category are closer to each other, while a few images are more distinct. The distributions are largely overlapped, but urban images have a smaller mean Euclidean distance (urban mean = 0.494) than natural images (nature mean = 0.606), and the differences in mean are significant in a two-sample t-test ($t = 88.627$, $df = 261680$, p-value < 0.001). After standardizing the twelve low-level features to uniform variances, however, the differences in mean between natural and urban

images are no longer significant (Figure 2b, t = -0.64146, df = 264693, p-value = 0.5212). In addition, the distributions of Euclidean distances of natural and urban images are largely overlapping, indicating that the Euclidean distance matrix in natural and urban images is not significantly different from each other. The same results also hold true when applying the cosine distance norm instead. In Figure 3a, the mean pairwise cosine distance for the urban image set is greater than that of the natural image set, and the difference is significant (t = 110.62, df = 250526, p-value < 0.001). After scaling, the difference in cosine distance norms is no longer significant either (Figure 3b, t = 0.016878, df = 264697, p-value = 0.9865). The non-significant result does not support our hypothesis that natural images are similar to each other in terms of low-level visual features, and the significant difference in average distance on unscaled low-level features data is potentially due to the unequal variances within the data and the large sample size.

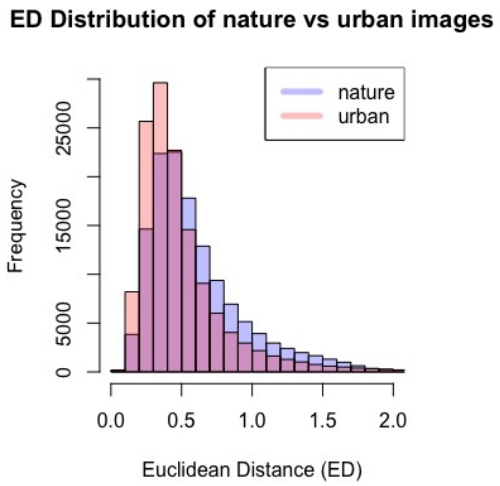

Figure 2a. Histogram of Euclidean distances of natural and urban images on unscaled low-level features

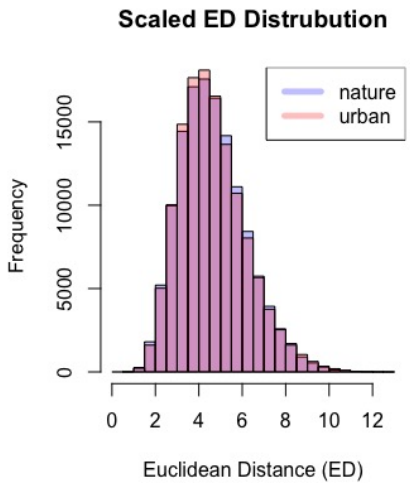

Figure 2b. Histogram of Euclidean distances of natural and urban images on scaled low-level features

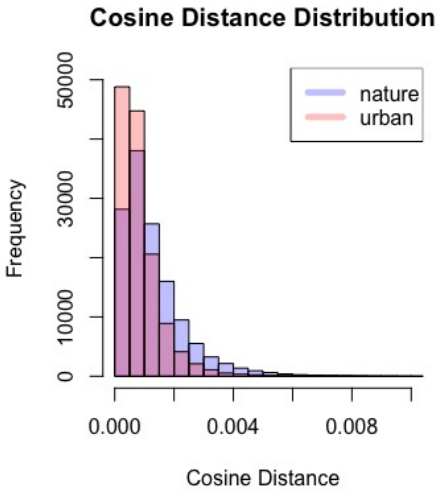

Figure 3a. Histogram of cosine distances of natural and urban images on unscaled low-level features

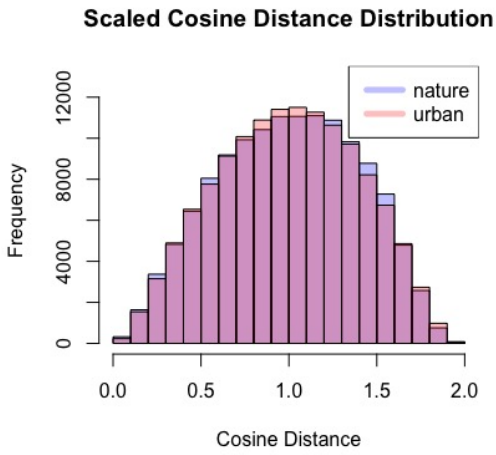

Figure 3b. Histogram of cosine distances of natural and urban images on scaled low-level features

Empirical Analysis 2: High-level Semantic Features

Empirical Analysis 2 focuses on the perceived similarities of images in high-level semantic features. High-level visual features refer to the information that the human brain processes after perceiving visual stimuli. Visual memory, object identification, and face recognition are several components of high-level visual processing. The present experiment focuses on the semantic information of images. To obtain semantic information of images, the present study applies the label detection algorithm from Google Vision API to the same image set as the low-level feature analysis to generate the semantic information for every image. Labels for each image are transformed into 300-dimensional vectors using the pre-trained word embedding model Word2Vec (Church, 2017), with a few exceptions if the word is not included in the pre-trained model. The corresponding confidence score for every semantic label is applied as the weight, and every image is represented by the average of weighted vectors. Similar to Empirical Analysis 1, the distances of every pair of

images within the natural and urban image set are calculated to compare the differences in perceived similarity. A small distance indicates high similarity and vice versa. However, only cosine distance is applied as the distance norm in this part of the analysis. Euclidean distance, on the other hand, is not a desirable choice in high-dimensional calculation due to the lack of variances in the distance norm.

Methods

Materials

The present experiment applies the label detection algorithm from Google Vision API on Google Cloud Platform to collect the semantic information of all valid images from Empirical Analysis 1. When given an image as input, the label detection algorithm provides a list of labels of the semantic theme that the algorithm detected with a confidence score above 50%. The default number of labels per image is 10, but fewer labels are possible if there are not enough labels passing the 50% confidence score threshold. Across all 1024 images, the label detection algorithm generates 10,236 labels. Figure 4 shows an example output from the Google Vision API label detection algorithm.

```
["'Cloud(97.50%)'",
 " 'Sky(97.23%)'",
 " 'Plant(96.25%)'",
 " 'Building(92.61%)'",
 " 'Leaf(89.35%)'",
 " 'Window(87.59%)'",
 " 'Land lot(85.97%)'",
 " 'Vegetation(84.89%)'",
 " 'Urban design(84.03%)'",
 " 'Residential area(80.66%)'"]
```

Figure 4. An example output from the Google Vision API label detection algorithm

There are three reasons why the present study uses the label detection

algorithm by Google Vision API instead of other object recognition algorithms: 1. The object identification algorithm focuses on the objects in an image and only provides the most likely results. In comparison, the label detection algorithm provides a list of all possible labels of the semantic theme with confidence scores, which better matches the semantic information when humans perceive an image. 2. After a few testing trials, the present study finds that the object recognition algorithm does not provide meaningful results when presented with an image without a prominent object, so it is not applicable to many wide-angle natural images. 3. A previous study used the label detection algorithm on Google Vision API to extract semantic information from natural and urban images and studied the correlation between semantic information and perceived naturalness (Hyam, 2017). In that study, the researcher defined the Calculated Semantic Naturalness (CSN) score as the ratio of natural labels among the total number of labels minus the ratio of urban labels among the total labels, and the CSN correlates with perceived naturalness ratings at a significant level. This study provided critical support to justify the application of Google Vision API by demonstrating its capability to represent the perceived naturalness of images. Thus, the credibility of using the label detection algorithm on Google Vision API to study image semantic information has been established.

Besides Google Vision API, the present study also uses a pre-trained Word2Vec model to transform semantic labels into vectors. The Word2Vec model (Church, 2017) is trained on part of the Google News dataset, which contains over 3 million words and phrases. The output of the Word2Vec model is a 300-dimensional vector. The present study adopts a pre-trained word embedding model instead of training a model based on the semantic labels generated because the present study aims to map the visual processing when humans perceive an image and reflect the

semantic information of that image. A word embedding model trained on billions of words would better match the retrieval process when humans search for a semantic label associated with an image. In addition, Word2Vec better fits the need of the present research question compared to other language processing models like BERT, which is more ideal for sentence transformation. The BERT model assigns different weights to words depending on their position within a sentence. The labels generated by the label detection algorithm are individual words and phrases. Taking all labels of an image and transforming them into vectors using sentence transformation rules would potentially bias the results because the labels may not construct a semantically meaningful sentence. Therefore, using Word2Vec to transform every label and take the weighted average of vectors would be a better choice for the present experiment.

To the best knowledge of the author of this paper, using the combination of Google Vision API with the pre-trained Word2Vec model is compatible with the purpose of the present paper. The workflow is such that the Google Vision API extracts the semantic information from images, and the pre-trained Word2Vec model transforms semantic labels into vectors and compares semantic similarities with high validity (Church, 2017). Although the credibility of this workflow has yet to be established by existing literature, there are a few advantages of using Word2Vec in comparison with other word embedding models (e.g., TFIDF, word frequency vectors, BERT) due to its large training size and validity in comparing word similarity.

Procedure

The steps to calculate the pairwise cosine distance within natural and urban image sets are as follows. To obtain the semantic labels, the present experiment passes all images with perceived naturalness ratings from empirical analysis 1 to the label detection algorithm. Then, all labels are transformed into 300-dimensional

vectors through the pre-trained Word2Vec model. The corresponding confidence score associated with the labels is applied as the weight, and the weighted average of vectors of all labels associated with an image is used to represent the image. For instance, when applying the confidence score as the weight, a label with a 90% confidence score is weighted as 0.9 when calculating the average of all labels. Similar to the methodology in low-level feature analysis, semantic feature analysis also calculates the distance matrices of natural and urban images. Since the image vectors have 300 dimensions, the semantic feature analysis uses cosine distance as the primary distance metric because of the lack of variances introduced by the Euclidean distance norm when applied to high-dimensional vectors.

Results

The Google Vision API label detection algorithm provides valid outputs for all 1024 images from empirical analysis 1 with no missing values. In total, 10236 labels are included. Most, except for 2 images, have 10 semantic labels (one image with 9 labels, one image with 6 labels) because the Vision API only reports labels with over 50% confidence score. Among all the labels, the Vision API provides 341 unique labels. After dividing the image dataset into natural and urban image sets by the median naturalness rating, the natural image set shares 200 unique labels, whereas the urban image set shares 242 unique labels. All labels that appear over 100 times are shown in Figure 5. The most frequent label across all images is "sky," with 795 of 1024 images containing this label. Among the 19 labels with over 100 occurrences, 13 of the 19 labels are natural labels (sky, plant, landscape, etc.), and only 6 of the 19 labels are urban labels (window, house, building, etc.). The disproportion of natural labels among the most frequent labels indicates that natural concepts are prevalent in the image dataset, suggesting that natural images potentially share more similar concepts than urban images.

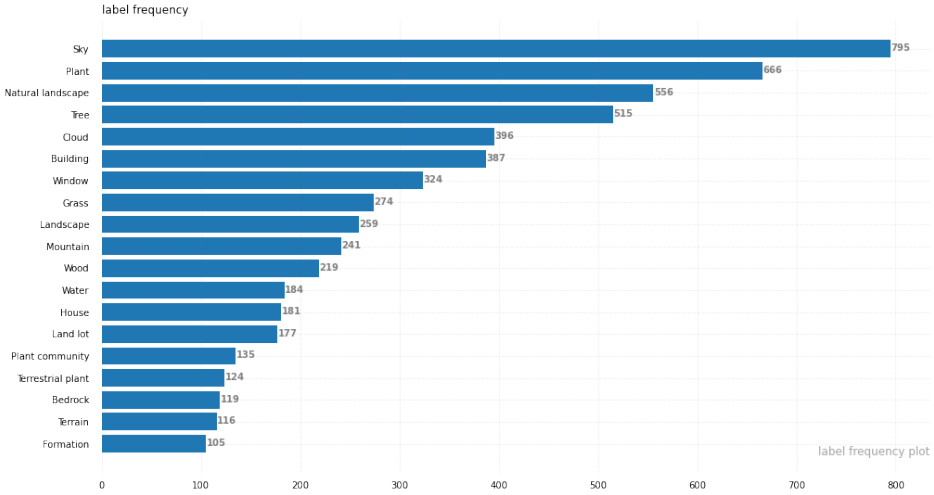

Figure 5. Label frequency count, showing labels that appear in over 100 images

To understand whether natural images share more similar semantic concepts than urban images, the present study calculates the cosine semantic similarity between images based on the labels of each image. The cosine similarity score ranges from 0 to 1, and a score close to 1 means high similarity and small distance between two images. The distribution of semantic similarity scores for natural and urban image sets are shown in Figures 6a and 6b. In both figures, the distributions of semantic similarity scores are slightly skewed to the left, and the average similarity score for natural images is 0.7093. In comparison, the average similarity score for urban images is 0.7286. The difference in average semantic similarity between natural and urban images is significant ($t = -58.71$, p-value < 0.001), meaning that the average semantic similarity in natural images is smaller. However, the significance in the difference in mean is potentially due to the large number of samples in each sample (n=262,144). When the sample size is large, a small difference in mean would lead to a statistically significant difference. To alleviate that bias, the average similarity score of each image to all other images within the set is also used to compare the difference in semantic similarity in a smaller sample size. In figure 7,

both the distributions of nature and urban images are left-skewed, whereas the distribution of nature images reaches the peak value at a smaller similarity score than that of urban images. The difference in semantic similarity score is also significant ($n = 512$, $t = -5.401$, $p < 0.001$), suggesting that urban images share similar semantic meaning to each other more than that of natural images, which does not support the original hypothesis.

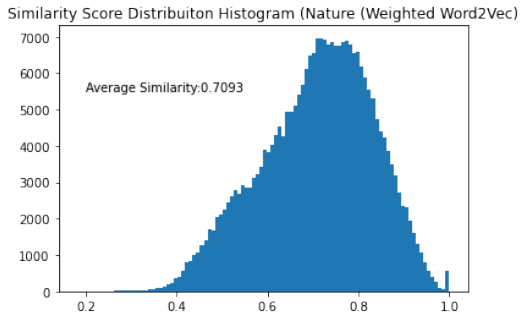

Figure 6a. distribution of pairwise cosine similarity score of natural image set

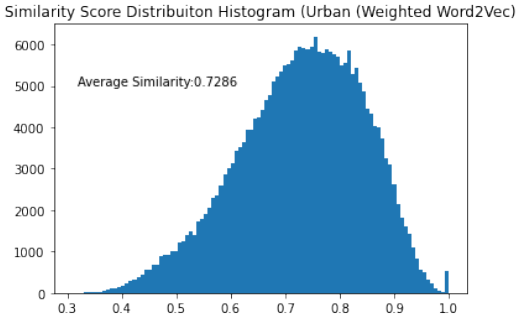

Figure 6b. distribution pairwise cosine similarity score of urban image set

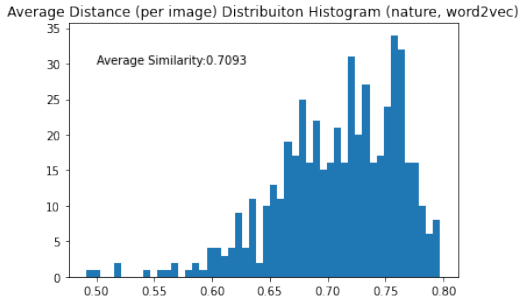

Figure 7a. distribution of average cosine similarity score of each natural image comparing to other natural images

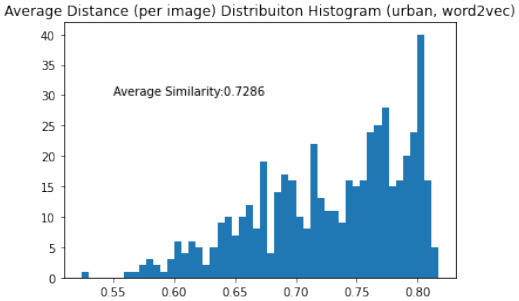

Figure 7b. distribution of average cosine similarity score of each natural image comparing to other natural images

In addition, the present study applies an alternative way of applying the confidence score of the semantic labels. Instead of using the confidence score as weights for the vectors from the Word2Vec model, the present study applies the confidence score as a threshold to filter the semantic labels. In this case, the semantic

representations of images are simplified, and the similarities between images are expected to decrease because of fewer overlapping semantic concepts. The present study chooses a 90% threshold to filter the labels, aiming to get a simplified but accurate semantic representation of the images. The pairwise cosine similarity scores are calculated in the same way as described above. As shown in figure 8 and 9, the similarity score distribution for the natural image set is separated into two clusters, with one cluster having relatively low similarity scores and the other having high similarity scores. This pattern is less obvious in the distribution for the urban image set, where most of the similarity scores are in the upper range. The difference in mean is significant (t = -92.83, p-value < 0.001), and the natural image set has a smaller similarity score on average. When looking at the distributions of the average similarity score per image, the urban image set has a larger number of images with high similarity scores. The difference is also statistically significant (t = -7.84, p-value < 0.001).

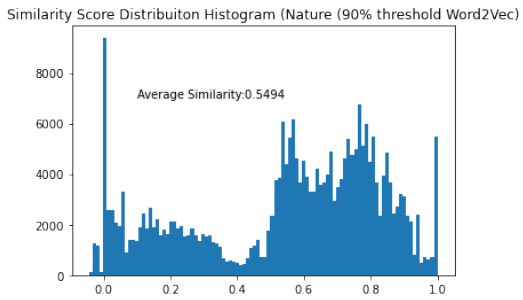

Figure 8a. distribution of pairwise cosine similarity score of natural image set when applying 90% confidence score threshold

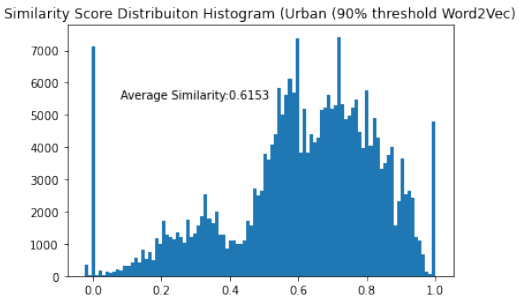

Figure 8b. distribution of pairwise cosine similarity score of urban image set when applying 90% confidence score threshold

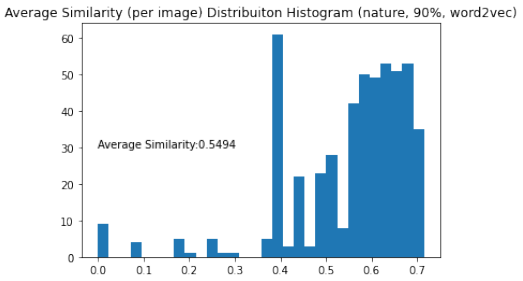

Figure 9a. distribution of average cosine similarity score of each natural image comparing to other natural images when applying 90% confidence score threshold

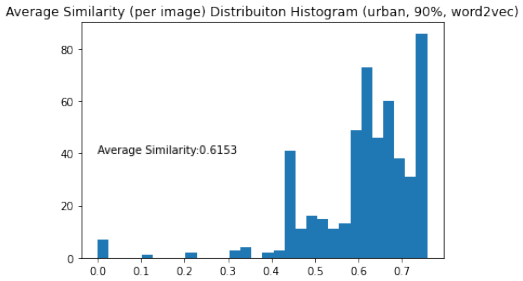

Figure 9b. distribution of average cosine similarity score of each urban image comparing to other urban images when applying 90% confidence score threshold

To validate the result, the present study also applied a modified One Hot Encoding technique instead of the pre-trained Word2Vec model to transform images into vectors. Specifically, the images are transformed into 341-dimensional vectors because there are 341 unique labels. Labels that match the images are marked as 1 in the corresponding position, and all other elements are marked as 0 in the vector. Instead of only one element among the 341 elements being coded as 1, this modified One Hot Encoding technique codes more than one element as 1 in the vector, as the number of elements matches the number of labels for each image. The pairwise cosine similarity scores are calculated for each image set as well.

As shown in the figure 10 and 11, the distributions of all similarity scores and the average similarity score per image for the natural image set and urban image set are similar. The difference in mean is significant when applying the large sample size ($t = -11.61$, p-value <0.001), but no longer significant when comparing the average similarity score per image ($t = -1.08$, p-value $= 0.28$), where the sample size is much smaller.

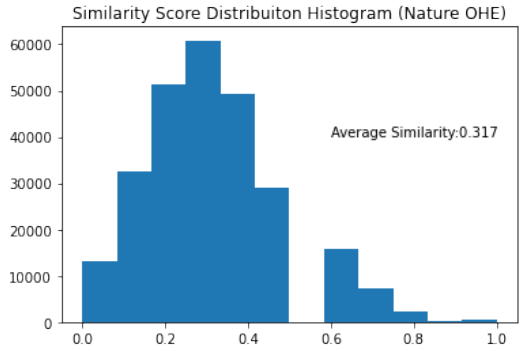

Figure 10a. distribution of pairwise cosine similarity score of natural image set when applying modified One Hot Encoding

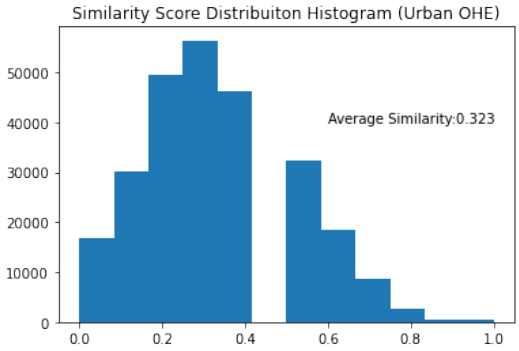

Figure 10b. distribution of pairwise cosine similarity score of urban image set when applying modified One Hot Encoding

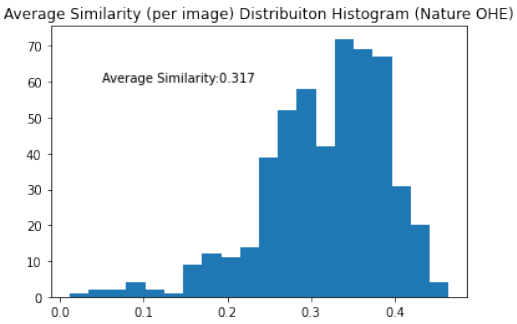

Figure 11a. distribution of average cosine similarity score of each natural image when applying modified One Hot Encoding

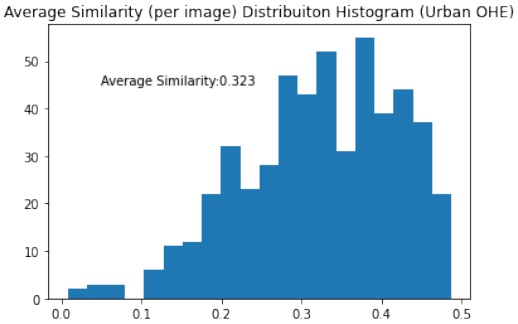

Figure 11b. distribution of average cosine similarity score of each urban image when applying modified One Hot Encoding

Besides that, the present study also coded the image vectors by semantic labels that have over a 90% confidence score to represent the images instead of using all semantic labels. As shown in the figure 12 and 13, the similarity score distributions are similar for natural and urban image sets. However, it is noteworthy that the natural image set has fewer entries with a 0 similarity score and more entries with a 1 similarity score than the urban image set. The 0 similarity score might be due to sparse vectors that only a few elements are non-zero values and the orthogonality between vectors, whereas the similarity score of 1 indicates more overlapping semantic labels in the natural image set, meaning the vector

representations are exactly the same. The difference in similarity score is significant (t=59.43, p-value <0.001), and the natural image set has a smaller mean similarity score than the urban image set. The average similarity score per image is also statistically significant (t=-5.033, p-value<0.001).

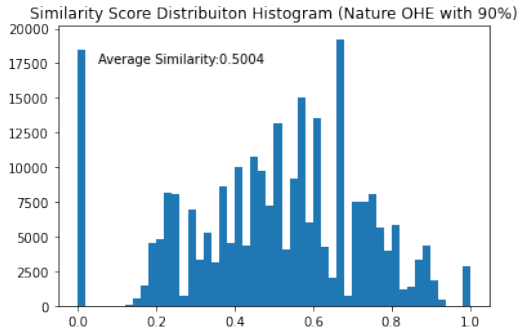

Figure 12a. distribution of pairwise cosine similarity score of natural image set when applying modified One Hot Encoding with 90% threshold

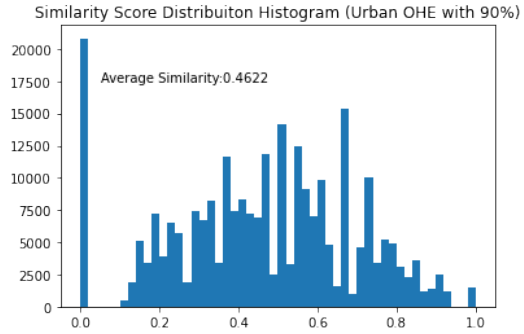

Figure 12b. distribution of pairwise cosine similarity score of urban image set when applying modified One Hot Encoding with 90% threshold

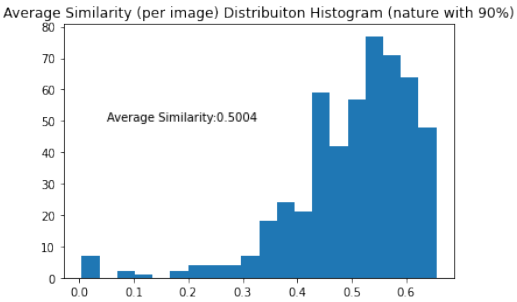

Figure 13a. distribution of average cosine similarity score of each natural image when applying modified One Hot Encoding with 90% threshold

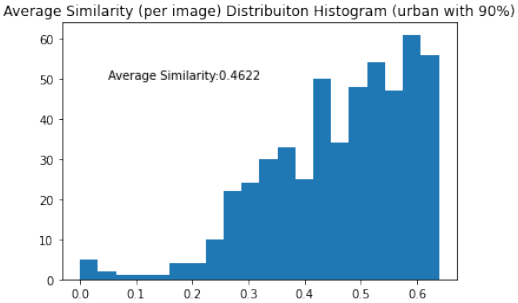

Figure 13b. distribution of average cosine similarity score of each urban image when applying modified One Hot Encoding with 90% threshold

To further analyze the differences, the present study also divides the image set into six subsets based on their naturalness ratings. As Figure 6b shows, the distributions of all image sets are different. The numbers of images are not evenly

distributed across the naturalness ratings. As shown in the figure, most of the images are either in the low naturalness rating group (naturalness rating between 1-2, n = 324) or high naturalness rating group (naturalness rating between 6 to 7, n = 430), whereas only 270 images are in the other groups. Thus, when applying the semantic similarity to each individual image subset and aligning the y-axis in the same figure, only group 1 and group 6 show meaningful results because of the unevenly distributed numbers of images per group.

Table 3: Naturalness Rating Counts

Naturalness Rating	Count
1-2	324
2-3	81
3-4	57
4-5	58
5-6	74
6-7	430

As the figure 14 shows, when applying a 90% confidence score threshold to filter the semantic labels per image, the similarity score distribution in the high naturalness rating group is separated into two clusters as well. Both the mean similarity scores for the low naturalness rating group (group 1, mean = 0.662) and the high naturalness rating group (group 6, mean = 0.564) are lower when compared to the undivided urban and natural image sets under the same coding mechanism. This result is aligned with our expectation because a narrower subgroup of images should be more similar to each other than a larger group. However, in the high naturalness rating group, some images are divergent from the other natural images,

driving the semantic similarity of natural images lower.

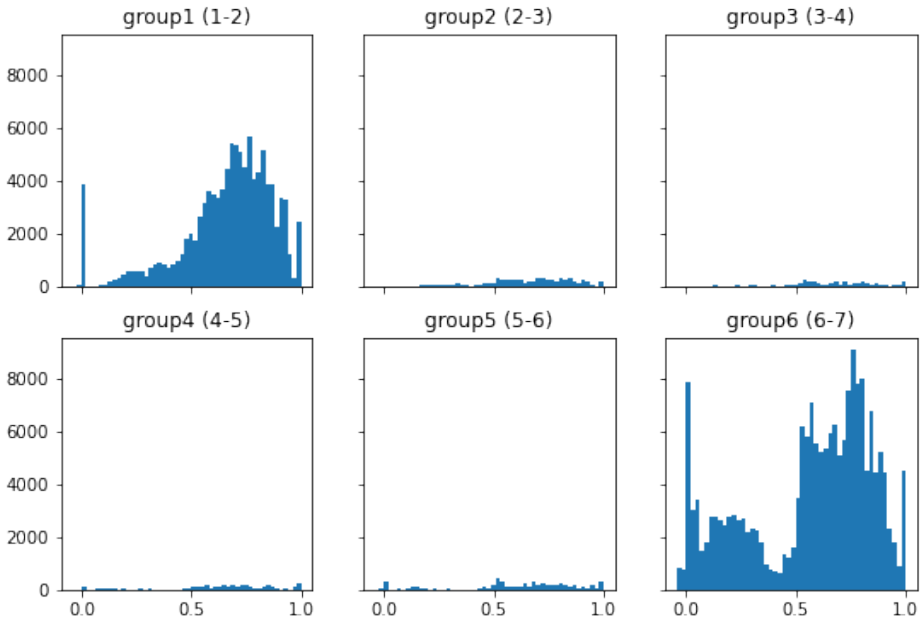

Figure 14. distribution of pairwise cosine similarity score of per image group

Discussion

To summarize, in the low-level feature analysis, the average pairwise distance in the natural image set is significantly smaller than that of the urban image set under both Euclidean and cosine distance norms before normalizing the variances. After normalizing the low-level features to equal variances, the distributions of pairwise distances overlap, and the difference in mean is no longer significant. Although the distinct properties of natural images contribute to a significant difference in average distances before normalizing the features, the significance might be attributed to unequal variances emphasizing some features rather than others when calculating the distance. However, it is not reasonable to neglect any visual features because all visual features contribute to an image collectively. Therefore, the findings fail to provide solid support for the assumption that natural images share

more similar low-level visual features than urban images.

For high-level semantic information, the hypothesis of the present study is that the natural image set shares a higher pairwise semantic similarity score (smaller distance) on average than the urban image set. If the hypothesis is correct, then the natural image set should have a significantly larger similarity score than the urban image set across all coding mechanisms. In the preliminary label frequency count, the number of unique labels generated from the natural image set is less than the number from the urban image set. Among the 19 labels that appear in at least 100 images, 13 of them are classified as natural, suggesting that natural images might share a narrower range of semantic concepts with more overlapping semantic words.

However, the semantic similarity analysis shows mixed results. When using all semantic labels weighted by confidence scores to represent the images, the average semantic similarity of the natural image set is significantly smaller than that of the urban image set. After downsampling the sample size by calculating the average similarity score per image, the significant difference in mean persists. When using semantic labels with over 90% confidence score to represent the images, the average similarity score of the natural image set remains significantly smaller. When further dividing the images into specific subsets, the distribution of the high naturalness rating set is separated into two clusters, with one cluster sharing relatively low similarity and the other sharing high similarity. The separated clusters suggest that the significant difference in similarity might be driven by a small set of natural images that are more divergent from each other, but further analysis is needed to understand how they are divergent.

On the other hand, contradicting results are found when applying modified

One Hot Encoding techniques to represent the images instead of the Word2Vec model. When using all the labels to calculate similarity scores, the difference in average similarity score per image is not significant. However, when only using labels with over 90% confidence scores, the average similarity score for the natural image set is significantly greater than that of the urban image set, both before and after calculating the average score per image. It is also noteworthy that the natural image set has fewer entries with a 0 similarity score and more entries with a 1 similarity score, which is compatible with the label frequency count result, indicating that labels from natural images are overlapping.

Overall, the present study does not provide strong evidence for the original hypothesis that natural images share similar low-level visual features and semantic concepts compared to urban images. While some results are aligned with our initial hypothesis, they do not hold up after normalization or under different coding techniques. The findings suggest that the relationship between memorability and similarity is likely more complex than a simple positive or negative correlation. Image memorability depends on a variety of factors that range from object atypicality (Bainbridge et al., 2013) to emotional valence(Khosla et al., 2015), and it is difficult to determine memorability from any single factor. A prior study focused on the relationship between image memorability and typicality and found that typicality cannot determine image memorability in some cases (Kramer et al., 2023). The study found that more memorable images tend to be more prototypical of their concepts in their representation, arguing against a general primacy of atypicality in memorability. However, when examining the relationship between memorability and object space typicality at the category level, the study found that certain object categories, such as containers and electronic devices, showed negative relationships, while other

categories like animals and body parts demonstrated positive relationships. Across all categories, the distribution of memorability-typicality relationships did not differ significantly from 0. This finding aligns with the results from the present study, suggesting that image memorability might be too complex to be explained by a simple positive or negative correlation of an individual factor. Another study has shown that even large sets of reasonably selected semantic attributes leave around 30% of unexplainable variance in predicting memorability of face images (Bainbridge et al., 2013). In the future, studies could focus on the intersectionality between two or more image properties (i.e., typicality-emotional valence) to unfold their relationship with memorability.

To further explore the variations and distinct properties of natural images, there are a few improvements that can be made to the present study. Firstly, the study only analyzed 1030 natural and urban images with perceived naturalness ratings, which is relatively small in the scope of computer vision research. However, the study's focus is different from computer vision research on image classification and categorization, as it aims to understand the perceived similarity of natural and urban images, which heavily relies on human perception. This is also why the study used human-rated perceived naturalness ratings instead of automatic classification ratings by algorithms, although algorithmic ratings would provide a larger sample size.

Secondly, the study is bounded by unknown algorithmic biases as it applies the Google Vision API and Word2Vec in high-level feature analysis. Although a previous study has shown the correlation between the perceived naturalness of images and their auto-generated semantic labels, it is unknown whether the Google Vision API provides accurate representations of semantic information in an image. A more accurate way to obtain semantic information from images is through human

experiments, but this would come at a higher cost compared to using automatic classification tools. Another model used in the study is the Word2Vec model trained on the Google news dataset. The study assumes that the pre-trained Word2Vec model can map human language processing due to its large training sample from the Google news database, but this is a bold assumption. The study finds that some image labels generated by the Google Vision API are not included in the pre-trained Word2Vec model, such as "Stalactite" and "Thatching." To resolve this, the study simply discards these words since they account for less than 3% of the total labels, but the impact on the semantic representation is unknown.

Furthermore, the results from the study are influenced by the models and methodologies being applied. To the best of the author's knowledge, two alternative models can be applied to re-examine the results. The first is DeepLabv3 (Chen et al., 2017), which is trained on the Cityscapes Dataset and employs atrous convolution to capture multi-scale context. DeepLabv3 is able to extract semantic features from the scene and generates semantic labels with the proportion of the object in the image, allowing for a better indicator to assign weights when transforming labels to vectors. However, the limitation of DeepLabv3 is that it is trained on street view images, and only twenty output labels are included. The present study did not choose the DeepLabv3 model due to its limited output categories, but proportion scores associated with labels would be desirable when doing label transformation.

The second model is a fine-grained deep ranking model that predicts image similarity with high accuracy. While the model is able to predict image similarity with high accuracy, the features used in the deep learning networks are not visible to users, and users cannot specify whether the two images are similar in low-level visual features or semantic features or both. However, the deep ranking model is a good

source to validate the image similarity results from the present study. In the future, researchers can employ these different label detection models, as well as human subject experiments, to strengthen the analysis of the present study.

Conclusion

The present study aimed to investigate whether natural images share more similar low-level visual features and high-level semantic concepts than urban images. The hypothesis was that natural images, being less memorable than urban images, would be more similar as a set. To test this hypothesis, the study transformed images into vectors based on low-level visual features and high-level semantic features, respectively. The study then calculated the Euclidean and cosine distance between every image pair in the low-level feature analysis and applied only cosine distance in the semantic similarity analysis. However, the study failed to provide strong evidence to support the hypothesis that natural images are more similar than urban images in either low-level or high-level dimensions. In the future, researchers could extend these findings by studying the intersectionality between multiple image properties to better understand their relationship with image memorability and cognitive benefits.

References

Bainbridge, W. A., Isola, P., & Oliva, A. (2013). The intrinsic memorability of face photographs. *Journal of Experimental Psychology: General*, *142*(4), 1323–1334. https://doi.org/10.1037/a0033872

Berman, M. G., Hout, M. C., Kardan, O., Hunter, M. R., Yourganov, G., Henderson, J. M., Hanayik, T., Karimi, H., & Jonides, J. (2014). The perception of naturalness correlates with low-level visual features of environmental scenes. *PLoS ONE*, *9*(12), e114572. https://doi.org/10.1371/journal.pone.0114572

Berman, M. G., Jonides, J., & Kaplan, S. (2008). The cognitive benefits of interacting with nature. *Psychological Science*, *19*(12), 1207–1212. https://doi.org/10.1111/j.1467-9280.2008.02225.x

Berman, M. G., Kross, E., Krpan, K. M., Askren, M. K., Burson, A., Deldin, P. J., et al. (2012). Interacting with nature improves cognition and affect for individuals with depression. *Journal of Affective Disorders*, *140*, 300–305. https://doi.org/10.1016/j.jad.2012.03.012

Capaldi, C. A., Passmore, H.-A., Nisbet, E. K., Zelenski, J. M., & Dopko, R. L. (2015). Flourishing in nature: A review of the benefits of connecting with nature and its application as a wellbeing intervention. *International Journal of Wellbeing*, *5*(4), 1–16. https://doi.org/10.5502/ijw.v5i4.449

Chen, L.-C., Papandreou, G., Schroff, F., & Adam, H. (2017). Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*.

Church, K. W. (2017). Word2vec. *Natural Language Engineering*, *23*(1), 155–162.

Greenwald, A. G., Nosek, B. A., Banaji, M. R., & Klauer, K. C. (2005). Validity of the salience asymmetry interpretation of the implicit association test: Comment on rothermund and wentura (2004). *Journal of Experimental Psychology: General*, *134*(3), 420–425. https://doi.org/10.1037/0096

Hiatt, L. M., & Trafton, J. G. (2013). The role of familiarity, priming and perception in similarity judgments. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *35*(7).

Hyam, R. (2017). Automated image sampling and classification can be used to explore perceived naturalness of urban spaces. *PLOS ONE*, *12*(1), e0169357. https://doi.org/10.1371/journal.pone.0169357

Kaplan, R., & Kaplan, S. (1989). *The experience of nature: A psychological perspective*. Cambridge University Press.

Kardan, O., Gozdyra, P., Misic, B., et al. (2015). Neighborhood greenspace and health in a large urban center. *Scientific Reports*, *5*, 11610. https://doi.org/10.1038/srep11610

Kellert, S. R., & Wilson, E. O. (1993). *The biophilia hypothesis*. Island Press.

Khosla, A., Bainbridge, W. A., Torralba, A., & Oliva, A. (2015). Understanding and predicting image memorability at a large scale. *Proceedings of the IEEE International Conference on Computer Vision*, 2390–2398. https://doi.org/10.1109/ICCV.2015.275

Klette, R., & Zamperoni, P. (1996). *Handbook of image processing operators*. Wiley.

Kotabe, H. P., Kardan, O., & Berman, M. G. (2016). Can the high-level semantics of a scene be preserved in the low-level visual features of that scene? a study of disorder and naturalness. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *38*(6).

Kotabe, H. P., Kardan, O., & Berman, M. G. (2017). The nature-disorder paradox: A perceptual study on how nature is disorderly yet aesthetically preferred. *Journal of Experimental Psychology: General*, *146*(8), 1126–1142. https://doi.org/10.1037/xge0000321

Kramer, M. A., Hebart, M. N., Baker, C. I., & Bainbridge, W. A. (2023). The features underlying the memorability of objects. *Science Advances*, *9*(17), eadd2981. https://doi.org/10.1126/sciadv.add2981

Mancas, M., & Le Meur, O. (2013). Memorability of natural scenes: The role of attention. *2013 IEEE International Conference on Image Processing*, 196–200. https://doi.org/10.1109/ICIP.2013.6738041

Menzel, C., & Reese, G. (2021). Implicit associations with nature and urban environments: Effects of lower-level processed image properties. *Frontiers in Psychology*, *12*, 591403. https://doi.org/10.3389/fpsyg.2021.591403

Nations, U. (2014). *World urbanization prospects: The 2014 revision.* http://esa.un.org/unpd/wup/Highlights/WUP2014-Highlights.pdf

Needell, C. D., & Bainbridge, W. A. (2022). Embracing new techniques in deep learning for estimating image memorability. *Computational Brain & Behavior*, *5*(2), 168–184.

Neumann, D., & Gegenfurtner, K. R. (2006). Image retrieval and perceptual similarity. *ACM Transactions on Applied Perception*, *3*(1), 31–47. https://doi.org/10.1145/1119766.1119769

Rim, N. W., Li, Y., Bainbridge, W. A., & Berman, M. G. (2023). *Are natural images less taxing than urban images? evidence from compressibility and memorability* [Manuscript in preparation].

Ruderman, D. L. (1994). The statistics of natural images. *Network: computation in neural systems*, *5*(4), 517. https://doi.org/10.1088/0954-898X_5_4_006

Schertz, K. E., Saxon, J., Cardenas-Iniguez, C., Bettencourt, L., Ding, Y., Hoffmann, H., & Berman, M. G. (2021). Neighborhood street activity and greenspace usage uniquely contribute to predicting crime. *npj Urban Sustainability*, *1*(1), 1–10. https://doi.org/10.1038/s42949-021-00019-9

Srivastava, A., Lee, A. B., & Simoncelli, E. P. (2003). On advances in statistical modeling of natural images. *Journal of the Optical Society of America A*, *20*(6), 1237–1250. https://doi.org/10.1364/JOSAA.20.001237

Torralba, A., & Oliva, A. (2003). Statistics of natural image categories. *Network: computation in neural systems*, *14*(3), 391. https://doi.org/10.1088/0954-898X_14_3_302

Ulrich, R. S., Simons, R. F., Losito, B. D., Fiorito, E., Miles, M. A., & Zelson, M. (1991). Stress recovery during exposure to natural and urban environments. *Journal of Environmental Psychology*, *11*(3), 201–230. https://doi.org/10.1016/S0272-4944(05)80184-7

Westerman, D. L., Lanska, M., & Olds, J. M. (2015). The effect of processing fluency on impressions of familiarity and liking. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(2), 426–438. https://doi.org/10.1037/a0038356

Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., & Torralba, A. (2010). Sun database: Large-scale scene recognition from abbey to zoo. *2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3485–3492. https://doi.org/10.1109/CVPR.2010.5539970