In the *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 2012

Small Sample Scene Categorization from Perceptual Relations

Ilan Kadar and Ohad Ben-Shahar
Dept. of Computer Science, Ben-Gurion University
Beer-Sheva, Israel
{ilankad,ben-shahar}@cs.bgu.ac.il

Abstract

This paper addresses the problem of scene categorization while arguing that better and more accurate results can be obtained by endowing the computational process with perceptual relations *between scene categories. We first describe a psychophysical paradigm that probes human scene categorization, extracts perceptual relations between scene categories, and suggests that these perceptual relations do not always conform the semantic structure between categories. We then incorporate the obtained perceptual findings into a computational classification scheme, which takes inter-class relationships into account to obtain better scene categorization regardless of the particular descriptors with which scenes are represented. We present such improved classification results using several popular descriptors, we discuss why the contribution of inter-class perceptual relations is particularly pronounced for under-sampled training sets, and we argue that this mechanism may explain the ability of the human visual system to perform well under similar conditions. Finally, we introduce an online experimental system for obtaining perceptual relations for large collections of scene categories.*

1. Introduction

The ability to categorize visual scenes rapidly and accurately is highly constructive for both biological and machine vision. Following the seminal demonstrations of the ability of humans to recognize scenes in a fraction of a second (e.g., [32, 5]), much research has been devoted to understanding its underlying visual process [26, 11, 27, 13, 21, 35, 2, 31, 43], which led to much interest in *computational* scene categorization as well. Although substantial progress has been made [29, 38, 27, 28, 41, 12, 44, 8, 18, 44, 30], the bulk of this visual process remains an open question, both behaviorally and computationally.

In our paper we follow previous attempts to define the (somewhat elusive) notion of "visual scene" to consider it as a semantically coherent, nameable view of a real-world environment in which humans can act [16, 27]. The *scene category* is defined as the basic-level category for a visual scene [40] and refers to the most common label used to describe it.

But what characterizes visual processing underlying this visual labeling process? In this work we focus on one possible aspect of this question related to prior knowledge about the *perceptual relations* between the different scene categories. To date, computational algorithms for scene categorization [29, 38, 27, 28, 41, 12, 44, 8, 18] rarely consider the possible effect of such perceptual relations. However, even intuitively, when our visual system observes a bedroom scene for a fraction of a second and "deliberates" how to categorize it, what possibly comes to mind in addition to "bedroom" are perhaps classes like "living room" or "kitchen". It appears as if our visual system does not even consider possibilities such as "coast" or "highway", or more generally, scenes which are perceptually "distant" from the observable reference class. Put differently, prior knowledge about the perceptual relations between the different categories of scenes may help facilitate better, more efficient, and faster categorization. As we argue later, knowledge of such relationships could partly explain the fact that humans are often able to learn and process scene categories from very few training examples while computational models usually need at least tens of training examples per-category before achieving reasonable categorization performance.

Exploring relations between categories is not new and has been previously explored in the context of object categorization in different forms. Several methods have been developed for building image hierarchy based on image features [15, 3, 1, 33, 34]. Griffin and Perona [15], for example, build a tree-like hierarchy of category relations by recursively splitting the set categories into two minimally confused subsets based on the confusion matrix that arise from the classier output. While the obtained relations seems to speed-up classification at a small cost of categorization performance, by construction they depend on the classifier and the selected features. Other methods proposed to incorporate both semantic and image features information in or-

der to build image hierarchies [19] or to transfer knowledge from large sample categories to under-sampled categories by sharing parts or features across categories [4, 10]. The use of semantic relations between categories was pushed even further by exploiting WordNet [23] as a semantic substrate for object recognition [22, 37, 14, 9]

Still, *semantic* relations between categories do not necessarily agree with their *perceptual* relationship. For example, concepts such as "snowy mountains" and "skiing activity" are far from each other semantically (e.g., see the Wordnet hierarchy [23]), while perceptually they are very much related. This observation is particularly salient in the context of visual scenes. For example, as we later show in Sec. 2, even our small experimental setup reveals that the "highway" category is perceptually closer to "coast" than to "kitchen", although semantically the opposite holds [23].

Acknowledging that perceptual relations between visual scene categories may have a central role in the categorization process, in this work we propose to infer these relations directly from human observers and then incorporate them in the computational model in a *way which is independent of the choice of descriptors and classifier*. In particular, our contributions and course of action is summarized as follows:

- We introduce an experimental paradigm for obtaining perceptual data about scene categorization (Sec. 2).
- We leverage the obtained insights to define perceptual relations between scene categories (Sec. 2.2).
- We extend a known non-parametric classifier (NBNN [7]) to a new algorithm that exploits inter-class relations (Sec. 3.2). We stress that similar extensions could be applied to other classifiers too.
- We combine our obtained perceptual relations between scene categories and our proposed classifier into a computational scene categorization framework that leads to significant improvements in scene categorization performance, especially when the number of training scenes in each category is small (Sec. 4).
- As a critical step toward the extension of our ideas, we introduce an online experimental system to establish perceptual relationships for large collections of scene categories (e.g., [44]) via participants from all over the world (Sec. 5).

2. Perceptual Relations Between Scene Categories

How could perceptual relations between scene categories be measured in a robust and unbiased fashion? Since, to our best knowledge, current models and empirical data do not address this issue, we propose a *category discrimination* paradigm where we briefly present two natural scene stimuli *simultaneously* and ask human observers whether they belong to the same scene category or not (i.e., same/different forced choice task). Doing so serves several goals: first, as we explain later, the results of such experiment could give us an empirical evaluation for the perceptual "distance" between scene categories. Second, by switching from *description*-based tasks (where subjects provide free form descriptions [11, 27]) or detection-based tasks (where subjects are required to confirm the observation of a given category [26, 25, 20, 21]), we remove any subjective bias both in subjects' response and in the interpretation of the data. Hence, we argue that a discrimination task is far more robust and reliable in terms of the results it can provide.

2.1. Experimental Setup and Dataset

The experiment was carried out in a dark room especially designed for psychophysics experiments, with 79 motivated undergraduate students having normal or corrected-to-normal vision. The dataset for the experiment consisted of 8 scene categories: coast, forest, mountain, highway, tall-buildings, street, kitchen and bedroom. The selection of categories was strongly influenced by various earlier studies [40, 11, 20], and consisted a range of categories as wide as 8 classes permits (e.g., natural scenes, manmade scenes, indoor scenes, and outdoor scenes). Scene images were borrowed from the corresponding categories of two published datasets [27, 12] and adapted in size to 256×256 pixels.

Fig. 1 depicts the sequence of events in an experimental trial. Each trial began with the simultaneous presentation of two images from our dataset for one of 6 different presentation times (PTs): 27, 40, 53, 80, 107, 1000 ms (all durations are multiplies of a 75Hz refresh cycle of a computer monitor and were selected randomly with equal probability). PTs were chosen to span a wide range, from very short up to a duration sufficient to allow elaborate perceptual description [11, 32, 6]. The longest PT was introduced as control for how well scene categories were defined semantically (since at 1000ms, categorization errors are unlikely attributed to perceptual confusion).

50% of the trials constituted a pair of images from the same category while the other 50% used images from different categories. Chance level performance was therefore 50%. After presentation for the selected PT, the two images were then masked by a pair of masks, each selected at random from a pool of eight random masks having $1/f$ amplitude spectrum [21]. The trial was concluded with a response cue which remained on screen until subjects' response. Participants pressed *Same* if they judged the two images to match in category or *Different* if not. They were encouraged to respond according to their first impression and as quickly and accurately as possible. Before beginning

the experiment, participants completed a category learning procedure where they viewed 9 images from each of the 8 participating categories so that they could get acquainted with the scene category labels. Subjects then completed 12 practice trials so they could become familiar with the experimental task and procedure, after which they started the 336 trial experiment itself. No learning or practice image was reused in the experiment. Experiments were self-paced and participants were allowed to take breaks any time. In practice, the 336 trial experiment took 20 min to complete on average.

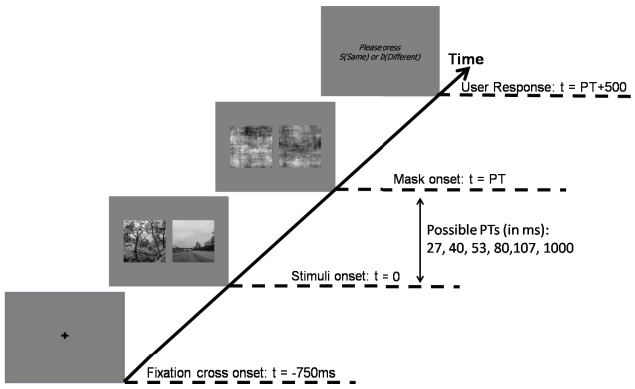

Figure 1. Experimental trials in our experiment begin with a fixation point, followed by a brief presentation of two images for PT ms and $1/f$ mask patterns for 500 ms. Subjects are then prompted to respond whether the two images belong to the same category or not.

2.2. Behavioral Results and Observations

How challenging is it to discriminate coast from forest scenes in a glance? How about a mountain and a bedroom scenes? Analyzing subjects' response first in the control trials with $PT = 1000ms$ reveals that discrimination under long stimulation yielded near-perfect performance of 96%, indicating that our scene categories are well-defined semantically. Eliminating this possible confound, we then explored the perceptual "distance" between all pairs of scene categories in our dataset by measuring subjects' accuracy for each pair over all trials and PTs (except the control trials). Table 1, which we term the *Perceptual Data (PD) Matrix*, shows the obtained accuracy, averaged over all PTs (except 1000ms), between all pairs of scene categories in our dataset. It is evident that subjects did not discriminate between categories equally accurately (or easily). For example, subjects were able to discriminate with much higher accuracy bedroom from forest (0.85), or mountain from street (0.84), but this accuracy dropped significantly when discriminating bedroom from kitchen (0.47), or forest from mountain (0.59), to name but a few.

Categories that are significantly less accurate to tell apart in the brief PTs used in our experiments must share enough perceptual properties to make the discrimination process more difficult (again, within these short PTs). In other words, Table 1 depicts the perceptual "*distance*" between all pairs categories for the purpose of scene categorization (e.g., with 0.85 compared to 0.68, the forest and kitchen are far *less* perceptually related than the highway and coast categories, respectively). Note that the results along the diagonal of Table 1 represent subjects' average accuracy in scenarios where the two presented images come from the same category, and hence represent a measure of "perceptual similarity" within each category, and not a perceptual distance. Clearly, at the level of category, the perceptual distance between a category to itself should be zero, and this is indeed how we define it in the sequel.

Although Table 1 represents the data in full, it may be useful to use it to organize the explored categories in a "perceptual space" in which the perceptual distance between class is more comprehensible. One way to carry out such visualization is *Multidimensional Scaling (MDS)* – a technique from statistical inference and data visualization to embed a set of objects in Euclidean space while preserving their "distance" as much as possible [36]. As can be appreciated from the results of such analysis in two dimensional space (Figure 2), the different classes that participated in our experiment appear to split into general two groups according to natural (left group) and manmade (right group) scenes. This intuitive division is also obtained formally once we apply the *k-means* clustering method to cluster the scenes to two clusters. The result, coded in color in Figure 2, shows a clear clustering along the natural/manmade classification, with one notable exception. Indeed, the highway category seems (at least marginally) closer to the natural scenes rather to its semantically related manmade scenes. Hence, even within this restricted 8-class experiment, we were able to find perceptual relations that do not conform to their semantic counterparts.

3. Small Sample Scene Categorization with Perceptual Relations

With perceptual relations established via proper experimental analysis as above (see Sec. 5 for discussion about handling larger collections of categories), we turn to discuss how they may be exploited for scene categorization, especially when only few labeled examples are available for each class. Indeed, in their recent attempt to provide a scene understanding database that "encompasses the richness and varieties of environmental scenes", Xiao *et al.* [44] reported that the majority of their 899 scene categories yielded only few image samples via Internet search. Consequently, their experiments excluded more than half of these categories and focused on those with at least 100 samples. Clearly, the requirement of well-sampled classes for satisfactory learning of scene categories may prove a critical obstacle vis-a-vis

	Bedroom	Kitchen	Street	Tall-building	Highway	Coast	Forest	Mountain
Bedroom	0.70	0.47	0.68	0.71	0.82	0.83	0.85	0.83
Kitchen	-	0.69	0.61	0.64	0.75	0.85	0.85	0.83
Street	-	-	0.74	0.58	0.55	0.81	0.82	0.84
Tall-building	-	-	-	0.81	0.82	0.83	0.81	0.82
Highway	-	-	-	-	0.62	0.68	0.80	0.78
Coast	-	-	-	-	-	0.60	0.74	0.68
Forest	-	-	-	-	-	-	0.71	0.59
Mountain	-	-	-	-	-	-	-	0.69

Table 1. The Perceptual Data Matrix obtained by measuring subjects' average accuracy between all pairs of scene categories in our dataset. The results along the diagonal represent subjects' average accuracy in scenarios where the two presented images come from the same category, and hence represent a measure of "perceptual similarity" within each category. All other entries represent the perceptual "distance" between their corresponding categories for the purpose of scene categorization.

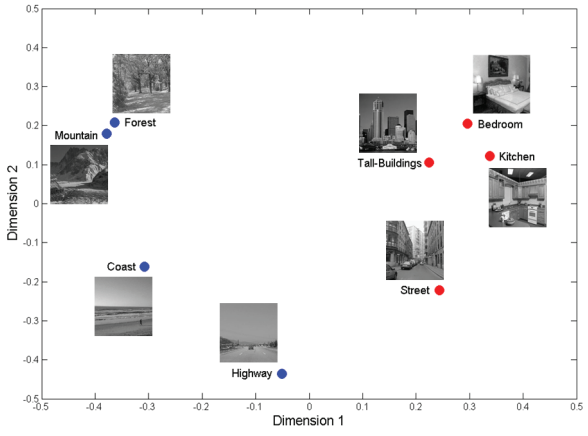

Figure 2. Applying MDS analysis on the perceptual distances obtained in our experiment, and visualizing the results as points in a 2D perceptual space, reveal a configuration which clusters natural scenes separately from manmade scenes (both intuitively and formally via k-means, the latter depicted with color). Interestingly, the clustering algorithm groups the highway category with the natural scenes, indicating that it is more perceptually related to the natural categories rather than to the manmade categories.

the frequency of under-sampled categories, a problem that is further increased by the high visual complexity of scenes, the rich and high dimensional representations that are typically used to capture that complexity, and the perceptual diversity within many known categories.

In contrast to existing computational algorithms, our everyday experience indicates that humans can learn new scene categories from only few examples, so it seems unlikely that a large set of training examples per-category is a necessity. In this paper we suggest that the solution resides in leveraging perceptual relations *between* categories in order to characterize each scene category in a more informative way.

To introduce the idea, consider that after learning few examples from each category, we are given a query scene and asked to decide whether or not it belongs to one of the categories, say "coast". Due to the visual complexity of real-scenes and the high visual variability within each scene category, it seems practically impossible that the few training coast examples would contain all the perceptual properties of a characteristic coast scene. However, additional perceptual properties of such scenes may possibly be obtained from training examples of other, perceptually related categories, e.g., the "highway" category (cf. the perceptual distance matrix in Table 1). Still, while doing so, highway examples should be considered as making smaller contribution compared to coast examples. More generally, we suggest that the perceptual properties of the given category can be learned or inferred not only from its designated exemplars, but also from all training examples that belong to other perceptually related categories, weighted according to their perceptual distance to the given category. In effect, such strategy increases the pool of useful training examples manyfold, thus facilitating categorization performance similar to well-sampled classes.

In what follows we develop this idea more formally and incorporate perceptual distance into a practical classifier. Although this can be done with almost any classifier, here we chose the *Naive-Bayes Nearest-Neighbor* (NBNN) algorithm [7] due to its excellent trade off between simplicity (or complexity) and performance. In particular, we extend the NBNN algorithm to incorporate inter-category relations and show how it provides a platform for coping with under-sampled categories.

3.1. Overview of NBNN

The NBNN algorithm due to Boiman *et al.* [7] is a non-parametric image classifier that employs 'Image-to-Class' (and not Image-to-Image) distances in the space of the local image descriptors. Additionally, it avoids descriptor quantization in order to cope with the large intra-category diversity within scene categories. Given a class $C_j, j \in \{1, 2, ..., t\}$ and query image Q with its corresponding local image descriptors $d_1, d_2, ...d_n$, the distance between Q and class C_j is defined as follows:

$$D(Q, C_j) = \sum_i NN_{C_j}(d_i) \quad (1)$$

where $NN_{C_j}(d_i)$ is approximation of $p(d_i|C_j)$ obtained by the distance between d_i and its nearest neighbor in class C_j, i.e.,

$$NN_{C_j}(d_i) = \min \left\{ d(d_i, d_k^{C_j}) \right\} , k \in \{1, 2, ..., L\} \quad (2)$$

where $d_1^{C_j}$,..,$d_L^{C_j}$ denote all the descriptors obtained from all the images contained in class C_j. Then, the classification of Q is done according the class C_j which minimizes $d(Q, C_j)$, i.e.,

$$\hat{C} = argmin_{C_j}(D(Q, C_j)) \ , \ j \in \{1, 2, ..., t\} \quad (3)$$

In their work, Boiman *et al.* [7] showed that the NBNN algorithm accurately approximates the theoretically optimal image classifier under the Naive-Bayes assumption (i.e., image descriptors are i.i.d given image class).

3.2. NBNN with Inter-Class Relations (NBNN-ICR)

Although the NBNN is quite effective for image classification, it treats each class independently and in isolation, and like most classifiers, it ignores possible relations between classes. But suppose that some relationship between classes is both meaningful and given, i.e., let $D_c(C_i, C_j)$ be a cost measure (or "distance") between classes C_i and C_j, as given by an outside source (e.g., Table 1). We assert that a simple extension to non parametric classifiers like NBNN can consider such information to enhance their performance. In particular, we suggest that searching for the nearest neighbor of d_i in class C_j should consider not only class C_j but *all* classes (i.e., $C_z, z \in 1, 2, ..., t$) while taking into account the distance measure $D_c(C_j, C_z)$. Formally, we suggest replacing Eq. 2 with the following computation

$$NN_{C_j}(d_i) = \min \left\{ d(d_i, d_k^{C_z}) + \alpha * D_c(C_j, C_z) \right\}, \quad k \in \{1, 2, ..., L\} \ ; \ z \in \{1, 2, ..., t\} \quad (4)$$

where $d_1^{C_z}$,..,$d_L^{C_z}$ denote all the descriptors obtained from all the images contained in class C_z and α is a weight parameter. Note that when $\alpha \to \infty$, one is unable to find the nearest neighbor $NN_{C_j}(d_i)$ of d_i in class other than C_j. Hence, NBNN can be considered a special case of NBNN-ICR.

What are the benefits of using Eq. 4 rather than Eq. 2? From a theoretical point of view, Eq. 2 will approximate $p(d_i|C_j)$ more accurately as the number of training examples per class approaches infinity [7]. However, as argued in the beginning of Sec. 3, in many practical cases, the number of training examples per class is very small compared to the class complexity, which renders the practical results by Eq. 2 depart significantly from the theoretical optimum. Eq. 4 attempts to alleviate this situation by exploiting relations between classes and employing information from *all* training examples, not necessarily from class C_j only. As we later show in 4, the fewer labeled class samples available, the more significant the improvement obtained by Eq. 4 over Eq. 2.

4. Experimental evaluation

4.1. Experimental Setup

Following the discussion above, we can now combine our measured perceptual relations and the NBNN-ICR classifier into a computational scene categorization framework. For that, we define the distance between any two scene classes C_i and C_j to be the measured distance from Table 1, i.e., we set $D_c(C_i, C_j) = PD(C_i, C_j)$. (Recall that by construction $PD(C_i, C_i) = 0$.) The resultant classifier, termed here as NBNN-PR, is then compared to NBNN to show how the use of the measured perceptual relations facilitates significant improvements in scene categorization performance, especially when the number of training scenes in each category is small. We also show that this improvement does not result from the mere inclusion of *any* class relations, but from the very particular relations that were inferred experimentally and reflect human perception. Toward this end we compared performance to instances of NBNN-ICR where the inter-class relations $D_c(C_i, C_j)$ for $i \neq j$ are selected randomly (abbreviated as NBNN-Rand), or are estimated computationally from a trained classifier confronted with the same experimental procedure as described for humans in Sec. 2.1 (abbreviated as NBNN-CR).

The dataset for the experiment consisted of the same scene categories described in Sec. 2.1: coast (360 images), forest (328 images), mountain (374 images), highway (260 images), tall-building (356 images), street (292 images), kitchen (151 images) and bedroom (177 images). In all cases we randomly split each category to disjoint training and testing sets, with $n_{training} = 1, 2, 4, 8, 16, 32, 64, 128$. The same sets where then used with the four algorithms (i.e., NBNN-PR, NBNN , NBNN-Rand, NBNN-CR) and repeated 20 times (to control for the random selection). Per-class categorization rates were then pooled across repetitions to produce average results.

4.2. Implementation

Since our approach is independent of the specific image descriptor used, the choice of the descriptor used is somewhat arbitrary for the purposes of this study. We selected two well known image descriptors that are potentially useful for scene classification (GIST and LBP) and also a trivial descriptor for baseline comparisons.

The GIST descriptor was proposed specifically for scene recognition tasks by Oliva and Torralba [27] and is based on the output energy of a bank of 32 Gabor-like filters tuned to 8 orientations at 4 different scales. Originally, the image re-

gion was divided to a 4×4 grid (e.g., 64×64 pixel patches for 256×256 image) where the squared output of each filter was averaged. However, to achieve greater discriminative power via higher feature dimensionality [7] we carried the same averaging operation on a 16×16 grid (i.e., 16×16 pixel patches for 256×256 image). Using the code provided by the authors [27] we therefore obtained 256 local descriptors of 32 dimensions for each given scene.

Since scene recognition may be regarded as texture classification [43], the *LBP* image-texture descriptor [24] was selected as our second descriptor. The LBP produces binary code for each pixel by comparing the pixel value to those values of pixels distributed uniformly around it. Then, a histogram measuring the frequency of each binary code is computed over a grid of non-overlapping regions. Again, using publically available code [44] we obtained 256 local descriptors (i.e., 16×16 grid) of 256 dimensions (i.e., 2^8 different binary patterns) for each given scene.

Thirdly, as a naive baseline, we also implemented a trivial image descriptor where the intensity values of the image are stacked and normalized over 16×16 patch grid to yield 256 local descriptors of 256 dimensions.

Since all variants of NBNN-ICR required setting value for α, a 2-fold cross-validation procedure was applied within the training set to select the *optimal* α for each run (except for runs with $n_{training} = 1$, where we selected the α that gave us the best performance over the test set). The search of α was restricted to the range $\alpha \in [0.05, 1]$ (in 0.05 step quantization) since an empirical test confirmed that $\alpha \approx 1$ is big enough to effectively nullify the effect of inter-class relations. We also note that the inter-class relations for the NBNN-CR classifier (see Sec. 4.1) were devised from a SVM classifier which was trained using the image descriptors defined above with 9 training images per-category.

4.3. Results

A performance comparison of NBNN, NBNN-PR, NBNN-Rand, and NBNN-CR is shown in Fig. 3 for all 3 descriptors. Table 2 shows the percentage improvement obtained by the NBNN-PR over the NBNN for the 3 different descriptors. As the results show, the use of the perceptual relations improves classification performance in all cases and all training set sizes, but is particularly significant when the number of training samples is relatively small. The graphs also show that the inter-class relations used cannot be arbitrary but rather they must reflect the true perceptual relations. Indeed, when fictitious relationships are used (as in NBNN-Rand1), the cross-validation procedure (which selects the best α) typically selected $\alpha \approx 1$ which effectively nullified the use of these relations and provided results similar to those of NBNN. Moreover, when forcing the use of these random relationships by using the same optimal α from NBNN-PR (as in NBNN-Rand2), performance of the NBNN-Rand has dropped severely even below NBNN, let alone compared to NBNN-PR. Finally, while using inter-class relations that are based on a particular choice of classifier and descriptor may improve performance compared to NBNN (cf. NBNN-CR), the use of *perceptual* relations which is independent of the choice of descriptor and classifier yields significantly better classification performance.

5. Learning Perceptual Relations for Many Scene Categories

One cannot avoid noticing that our experimental evaluation was limited to 8 scene categories, a fact that at first sight is inconsistent with the availability of much larger collections [44] of the standards that the image classification community has become used to, especially for object recognition [14, 15]. The reason for this is not computational, however. Rather, it is unfortunately very difficult to obtain reliable perceptual data for many scene categories while using controlled lab procedures like those described in Sec. 2. There are two main reasons for this. First, since subjects need to remember the scene categories that participate in the experiment in order to facilitate their "same" or "different" decision, they fail to do so when the number of categories exceeds some memory threshold. Second, the number of trials in such an experiment grows quadratically with the number of participating categories, and reaches quickly the capacity of normal human subject. A pilot experiment with 15 categories already exhibited both of these problems and prevented us from presenting corresponding results.

As a result of the above, it is clear that the data collection procedure must be amended to facilitate the acquisition of perceptual relationships between many categories, with the ambitious goal may be one that targets the SUN database of 899 categories [44]. One way to achieve this is by repeating numerous times the experiment from Sec. 2, each time for a small number of categories selected randomly from the large database. Pooling data from a huge number of participants may then overcome both limitations and provide the knowledge base to construct the perceptual relations between all categories in the database. Toward the goal above, we have developed an online experimental system that harnesses the power of the web to facilitate such data collection. Relying on the collective effort of a large population of users had already proved successful (e.g., [42, 39]) and we hope it could prove constructive in our case also.

Our web system is a *Silverlight* application that allows users from all over the world to participate in our experiment [17]. This application executes an experiment identical to the one described in Sec. 4.1, but in a form of a game that motivates users to participate in it.

In the beginning of each experiment participants are shown the instructions while the system randomly selects

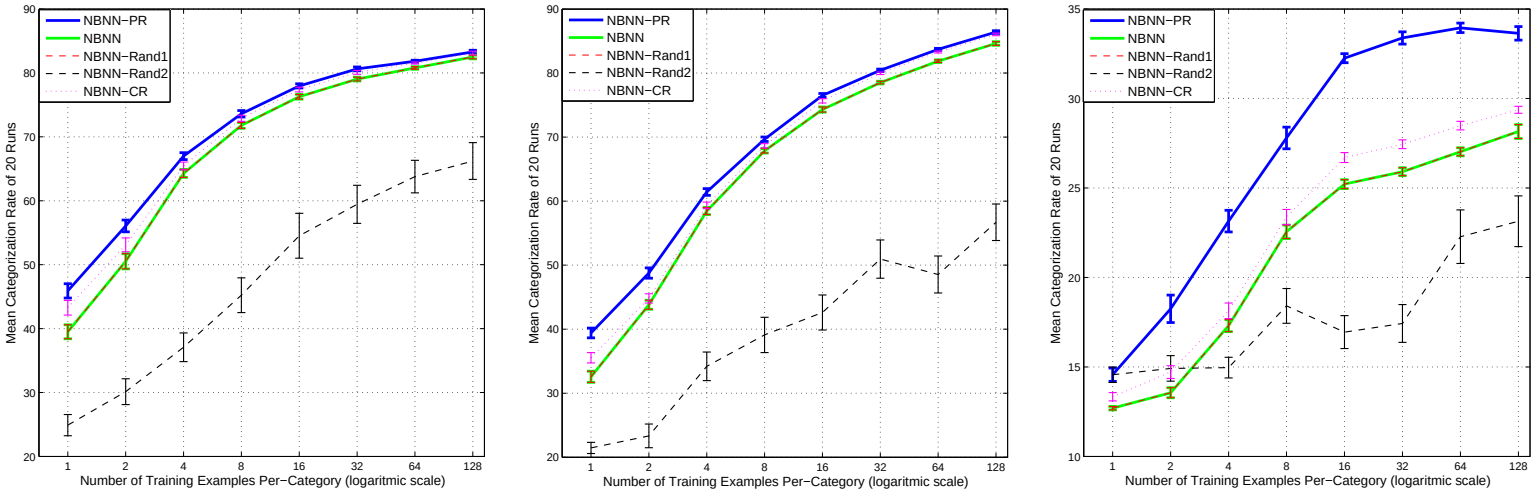

Figure 3. Performance of all discussed classifiers (NBNN, NBNN-PR, and all other variants of NBNN-ICR), based on all three tested descriptors, as the number of training examples is increased.

Descriptor	Number of training examples per category							
	1	2	4	8	16	32	64	128
GIST	16.2	11	4.2	2.5	2.2	2	1.3	1
LBP	21	11.3	5.1	2.7	3	2.4	2.3	2.1
Intensity	14.7	35	33.8	23.3	27.9	28.9	25.6	19.5

Table 2. Performance improvement (in percentages) obtained by NBNN-PR over NBNN for the three different descriptors tested, all as a function of training set size. The performance improvement is calculated as the difference between the average performance of the NBNN-PR and the NBNN, divided by the value of the latter.

few different categories out of the many categories of the SUN database [44]. Participants then need to complete a category learning procedure and a short practice session so that they could get acquainted with the scene category labels and the experimental procedure and task. The real experiment follows these steps and consists trials of the form discussed in Sec. 2.

Unlike in the lab, using web application suffers from the inability to control several experimental parameters, the most critical of which are presentation times. Very short presentation times are excluded for inability to ensure small relative error in their value when executed on unknown computer platform and display device. Hence, we currently limit PTs to 50, 100, and 200 ms, using the latter as "catch trials" to validate subject's awareness (High error rates in this PT would indicate unreliable subject). Except as noted, the sequence of events in an experimental trial are identically to those of 4.1.

With this online experimental system, we believe that the perceptual relations between many categories in the SUN database [44] could be established in reasonable period, and we call upon the community to participate in this experiment.

6. Summary and Future Work

In this work we argue that prior knowledge about the perceptual relations between different scene categories may help facilitating better and more accurate computational framework for the purpose of scene categorization. We first introduced an experimental procedure whose goal is to gain new insights on the visual process underlying human scene categorization. Then, we leverage the obtained insights to define perceptual relations between scene categories. While such inter-class relations could benefit many classifiers, we introduce an extension to the NBNN classifier and show how it significantly improves its performance, especially when supervised categories are under-sampled. Furthermore, we show that this improvement does not result from the mere inclusion of arbitrary inter-class relations but rather from from the very particular relations that were inferred experimentally and reflect human perception. Finally, seeking to apply this theory on large collections of scene categories, we introduce an online experimental system for mass acquisition of perceptual relations. We call upon the community to help collecting this data and we hope it would facilitate more accurate computational frameworks for learning hundreds of scene categories from only few-examples.

Acknowledgments

This work was funded in part by the European Commission in the 7th Framework Programme (CROPS GA no. 246252), the Frankel fund, the Paul Ivanier center for Robotics Research, and the Zlotowski Center for Neuroscience at Ben-Gurion University.

References

[1] N. Ahuja and S. Todorovic. Learning the taxonomy and models of categories present in arbitrary images. In *Proceedings of the IEEE International Conference on Computer Vision*, 2007. 1

[2] N. BaconMace, M. Mace, F. Thorpe, and T. S.J. The time course of visual processing: Backward masking and natural scene categorisation. *Vision Research*, 45:1459–1469, 2005. 1

[3] E. Bart, I. Porteous, P. Perona, and M. Welling. Unsupervised learning of visual taxonomies. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2008. 1

[4] E. Bart and S. Ullman. Cross-generalization: learning novel classes from a single example by feature replacement. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2005. 2

[5] I. Biederman. Perceiving real-world scenes. *Science*, 177:77–80, 1972. 1

[6] I. Biederman. On the semantics of a glance at a scene. In M. Kubovy and J. Pomerantz, editors, *Perceptual Organization*, pages 213–254. Lawrence Erlbaum, 1981. 2

[7] O. Boiman, E. Shechtman, and M. Irani. In defense of nearest-neighbor based image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2008. 2, 4, 5, 6

[8] A. Bosch, A. Zisserman, and X. Munoz. Scene classification via pLSA. In *Proceedings of the European Conference on Computer Vision*, pages 517–530, 2006. 1

[9] T. Deselaers and V. Ferrari. Visual and semantic similarity in imagenet. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1777–1784, 2011. 2

[10] L. Fei-Fei, R. Fergus, and P. Perona. One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28:594–611, 2006. 2

[11] L. Fei-Fei, C. Koch, A. Iyer, and P. Perona. What do we see when we glance at a scene? *Journal of Vision*, 7:1–29, 2007. 1, 2

[12] L. Fei-Fei and P. Perona. A bayesian hierarchy model for learning natural scene categories. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2005. 1, 2

[13] L. Fei-Fei, R. VanRullen, C. Koch, and P. Perona. Rapid natural scene categorization in the near absence of attention. *Proceedings of the National Academy of Sciences of the USA*, 99:9596–9601, 2002. 1

[14] R. Fergus, H. Bernal, Y. Weiss, and A. Torralba. Semantic label sharing for learning with many categories. In *Proceedings of the European Conference on Computer Vision*, 2010. 2, 6

[15] G. Griffin and P. Perona. Learning and using taxonomies for fast visual categorization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2008. 1, 6

[16] J. Henderson and A. Hollingworth. High level scene perception. *Psychological Review*, 50:243–271, 1999. 1

[17] I. Kadar and O. Ben-Shahar. An online experimental system: `http://www.cs.bgu.ac.il/~vision/pmc`, 2011. 6

[18] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2169–2178, 2006. 1

[19] L. Li, C. Wang, Y. Lim, D. Blei, and L. Fei-Fei. Building and using a semantivisual image hierarchy. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2010. 2

[20] L. Loschky and A. Larson. The natural/man-made distinction is made before basic-level distinctions in scene gist processing. *Visual Cognition*, 18:513–536, 2010. 2

[21] L. Loschky, A. Sethi, and D. Simons. The importance of information localization in scene gist recognition. *Journal of Experimental Psychology: Human Perception and Performance*, 33:1431–1450, 2007. 1, 2

[22] M. Marszalek and C. Schmid. Semantic hierarchies for visual object recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2007. 2

[23] G. Miller. Wordnet: A lexical database for english. In *Communications of the ACM*, 1995. 2

[24] T. Ojala, M. Pietikainen, and D. Harwood. Performance evaluation of texture measures with classification based on kullback discrimination of distributions. In *Proceedings of the IEEE International Conference on Pattern Recognition*, pages 582–585, 1994. 6

[25] A. Oliva. Gist of the scene. In L. Itti, G. Rees, and J. Tsotsos, editors, *Neurobiology of Attention*, pages 251–256. Elsevier, 2005. 2

[26] A. Oliva and P. Schyns. Coarse blobs or fine edges? evidence that information diagnosticity changes the perception of complex visual stimuli. *Cognitive Psychology*, 34:72–107, 1994. 1, 2

[27] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001. 1, 2, 5, 6

[28] A. Oliva and A. Torralba. Depth estimation from image structure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24:1226–1238, 2002. 1

[29] A. Oliva, A. Torralba, A. Guerin-Dugue, and J. Herault. Global semantic classification of scenes using power spectrum templates. In *Challenge of Image Retrieval*, 1999. 1

[30] D. Parikh and K. Grauman. Relative attributes. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 503–510, 2011. 1

[31] M. V. Peelen, L. Fei-Fei, and S. Kastner. Neural mechanisms of rapid natural scene categorization in human visual cortex. *Nature*, 2009. 1

[32] M. Potter and E. Levi. Recognition memory for a rapid sequence of pictures. *Journal of Experimental Psychology*, 81:10–15, 1969. 1, 2

[33] C. Schmid. Constructing category hierarchies for visual recognition. In *Proceedings of the European Conference on Computer Vision*, 2008. 1

[34] J. Sivic, B. Russell, A. Zisserman, W. Freeman, and A. Efros. Unsupervised discovery of visual object class hierarchies. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2008. 1

[35] S. Thorpe, D. Fize, and C. Marlot. Speed of processing in the human visual system. *Nature*, 381:520–522, 1996. 1

[36] W. S. Torgerson. Multidimensional scaling: theory and method. *Psychometrika*, 17(6):401–419, 1952. 3

[37] A. Torralba, R. Fergus, and F. W.T. 80 million tiny images: a large dataset for non-parametric object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(11):1958–1970, 2008. 2

[38] A. Torralba and A. Oliva. Statistics of natural images categories. *Network: Computation in Neural Systems*, 14:391–412, 2003. 1

[39] A. Torralba, B. C. Russell, and J. Yuen. Labelme: online image annotation and applications. In *Proceedings of the IEEE*, pages 1467–1484, 2010. 6

[40] B. Tversky and K. Hemenway. Categories of environmental scenes. *Cognitive Psychology*, 15:121–149, 1983. 1, 2

[41] J. Vogel and B. Schiele. Semantic typicality measure for natural scene categorization. In *Annual Pattern Recognition Symposium*, 2004. 1

[42] L. Von Ahn and L. Dabbish. Labeling images with a computer game. In *Human Factors in Computing Systems*, pages 319–326, 2004. 6

[43] L. Walker and J. Malik. When is scene recognition just texture recognition? *Vision Research*, 44:2301–2311, 2002. 1, 6

[44] J. Xiao, J. Hays, K. Ehinger, A. Oliva, and A. Torralba. Sun database: Large scale scene recognition from abbey to zoo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2010. 1, 2, 3, 6, 7