In the *Workshop on Computer vision + ONTology Applied Cross-disciplinary Technologies (CONTACT)*, ECCV 2014

SceneNet: A Perceptual Ontology for Scene Understanding

Ilan Kadar and Ohad Ben-Shahar

Ben-Gurion University of the Negev

Abstract. Scene recognition systems which attempt to deal with a large number of scene categories currently lack proper knowledge about the perceptual ontology of scene categories and would enjoy significant advantage from a perceptually meaningful scene representation. In this work we perform a large-scale human study to create "SceneNet", an online ontology database for scene understanding that organizes scene categories according to their perceptual relationships. This perceptual ontology suggests that perceptual relationships do not always conform the semantic structure between categories, and it entails a lower dimensional perceptual space with "perceptually meaningful" Euclidean distance, where each embedded category is represented by a single prototype. Using the SceneNet ontology and database we derive a computational scheme for learning non-linear mapping of scene images into the perceptual space, where each scene image is closest to its category prototype than to any other prototype by a large margin. Then, we demonstrate how this approach facilitates improvements in large-scale scene categorization over state-of-the-art methods and existing semantic ontologies, and how it reveals novel perceptual findings about the discriminative power of visual attributes and the typicality of scenes.

Keywords: Scene understanding, scene gist recognition, scene categories, perceptual relations, perceptual space

1 Introduction

Scene recognition is a challenging problem in high-level computational vision, especially when the number of categories is large. While humans are able to learn and process hundreds of scene categories, the performance of existing scene recognition approaches drops dramatically as the number of categories increases [1]. In this paper we address two important limitations in the development of scene recognition systems which deal with a large number of categories: (a) the lack of proper knowledge about the ontology of scene categories; and (b) the absence of meaningful scene representation. To address both points, we introduce a new ontology database called "SceneNet" [2], a comprehensive ontology of scene categories that was derived directly from human vision via a large-scale human study. The SceneNet ontology organizes scene categories according to their *perceptual* relationships and provides lower dimensional scene representation with "perceptually meaningful" (Euclidean) distance measure, all of which facilitate large-scale scene understanding operations.

While the concept of SceneNet is general, in this paper we report of SceneNet-100, the current version which consists of 100 scene categories from the SUN database [1], with the eventual goal of targeting all of its 908 categories. as we demonstrate later, in addition to significantly better computational results on various large-scale scene understanding operations, the SceneNet database provides important insights into human scene representation and organization and may serve as a key element in better understanding of this important perceptual capacity.

While traditional scene recognition approaches rarely consider the possibility of ontological organization of scene categories and indeed treat each category separately and independently [3, 1, 4, 5], learning and using ontologies of categories is not new and has been explored in the context of object recognition in different forms in the past [6–17]. For example, several approaches have been developed for learning ontologies based on image features [6–10] to speed up classification for a small cost of categorization performance. However, by construction these approaches depend on the classifier and the selected features. The use of ontologies was recently promoted by exploiting WordNet [18] as a semantic relationships database for object recognition [12, 14–16, 19, 17]. For example, researchers have shown the benefits of using WordNet for organizing images [16], reducing computational complexity [12], improving classification and search engine results [14, 17], and learning similarity functions for better image retrieval [19]. Indeed, semantic relationships can be extracted quite conveniently from WordNet. Still, as we will show later in Sec. 2, *semantic* relationships between categories do not necessarily agree with their *perceptual* relationships.

Arguing that a proper knowledge of the ontology of scene categories should be based on perceptual criteria and inferred from human vision, our contributions and course of action are summarized as follows:

- We perform a large-scale human study to create the SceneNet-100 database, a publically available online ontology for scene understanding that organizes scene categories according to their perceptual relationships.
- We embed scene categories along with their perceptual relationships in a lower dimensional *perceptual space* with "perceptually meaningful" Euclidean distance, where each category is represented by a single prototype.
- We extend the large margin taxonomy embedding algorithm [20] to kernels for learning a non-linear mapping of scene images into the perceptual space, where each scene image is closest to its category prototype than to any other prototype by a large margin.
- We show how our approach leads to significant improvements in large-scale scene categorization over state-of-the-art methods and existing semantic ontologies.
- We exploit the proposed SceneNet database for novel perceptual findings about the discriminative power of visual attributes and the typicality of scenes.

2 SceneNet: An Online Database for Scene Understanding

Establishing a comprehensive ontology of real-world scenes is critical for further research in scene understanding. In this section we describe the construction of our large-scale *perceptual* ontology derived directly from human vision. To this end, we first perform a large-scale human study to determine the perceptual relationships between scene categories using a large set of scene categories that approximates the richness of the real world. Next, we embed the scene categories in a lower dimensional *perceptual space* which represents the perceptual relationships between classes in a meaningful and usable manner.

2.1 The Scene Categorization Pair-Matching Task

In order to measure the perceptual relationships between scene categories, we develop a crowd-source version of the "category discrimination task" recently proposed by Kadar and Ben-Shahar [21]. In particular, we presented workers on Amazon Mechanical Turk (AMT) with a *Scene Categorization Pair-Matching Game*, where participants viewed a series of *briefly* presented pairs of scenes and were asked to judge whether the two scenes belong to the same category or not (i.e., same/different forced choice task). Given the short exposure times (see below), this seems a rather challenging task. Still, evidence for parallel processing in high level categorization of natural images has already been reported, showing that humans are as fast in dual scene presentations as they are for single scene presentations [22].

The dataset for our "game" consisted of 100 scene categories borrowed from the SUN database [1]. The selection of scene categories was carefully done to focus on categories that represent minimal semantic confusion and are maximally diverse and representative of the space of scenes. Scene images were reduced to monochrome and adapted in size to 256×256 pixels.

Each trial of the experiment began with a fixation cross, followed by the simultaneous presentation of two images from our dataset for one of 3 different presentation times (PTs): 50, 100, 200 ms (Note that PTs shorter than 50ms were excluded for inability to ensure small relative error in their value when executed on unknown computer platform and display device via AMT). The longest PT was introduced as control (i.e., "catch trials") to verify participants' awareness. High error rates in this PT would indicate unreliable participant (see below). By design, 50% of trials in our experiment constitutes 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. Participants pressed *Same* if they judged the two images to match in category or *Different* if not.

In the beginning of each experiment participants were shown the instructions while the system randomly selected 4 different categories out of the total 100.

Participants then completed a category familiarity procedure using 24 images (6 from each category) so that they could get acquainted with the scene category labels. Then they ran 5 practice trials so they could become familiar with the experimental procedure and task. The experiment itself followed all these steps and consists of 50 trials. Including category familiarity and practice phases, the entire experiment lasted around 5 minutes for each participant.

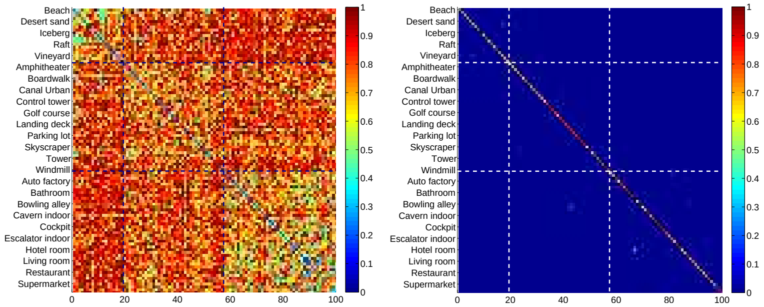
A total of **3262** workers from AMT (with better than 96% approval rate and at least 500 approved HITs) performed the game to provide a large pool of **163,100** trials. Workers were compensated with $0.5 per HIT, plus $0.1 bonus to high-scoring participants ($>$ 85% average discrimination accuracy in the experiment itself) to motivate them to do their best.

The primary difficulty of using a large, non-expert work-force is ensuring that the collected data is reliable. We first analyzed participants response in the catch trials with PT=200ms to confirm participants awareness. To exclude unreliable participants, we set a threshold = 0.75% on average discrimination accuracy (i.e., at the mid point between chance level and perfect discrimination) in PT=200ms trials (and only these trials). Once unreliable participants were filtered out, we were left with **2229** reliable participants over all PTs, whose response data was then used for the analysis and construction of our database.

2.2 Building Perceptual Ontology

Having all (reliable) subjects' response in the same/different discrimination task, we then explored the perceptual relationships between all pairs of scene categories in our dataset by analyzing discrimination accuracy over all trials and PTs. In particular, we calculated subjects' probability to respond *Different* for each pair of categories. Since this probability is expected to increase when such judgment is easier, and since the latter case is expected when scenes become more "perceptually different", this probability is termed as the "perceptual distance" (PD) between pair of visual scene categories [21]. But what are the benefits that such information may provide? We compare the matrix of perceptual distance (PD) with SUN's human confusion matrix [1]. A visual depiction of the two matrices is shown in Fig. 1. Both are organized according to the main semantic classes of the SUN's manually defined ontology [1] (Natural, Manmade Outdoor, and Indoor categories). Fig. 2 further illustrates the perceptual distance with several examples.

Several conspicuous initial observations can be made upon inspection of Fig. 1. First, while the vast majority of entries in the SUN confusion matrix are zeros, the entries in the PD matrix varies between all pairs of scene categories in our dataset to obtain a more informative matrix that can be used for building ontology. Second, given the unlimited presentation time, the confusions in the SUN confusion matrix are likely to be semantic-based rather than perceptual-based (e.g., SUN workers confused between Canal-Urban and Canal-Natural while perceptually they are far apart with PD=0.77; at the same time SUN workers did not confuse Beach and Desert-Sand while perceptually they share similar perceptual properties, PD=0.42). Indeed, quite often the perceptual relationships



(a) Perceptual Distance Matrix (b) SUN Matrix [1]

Fig. 1. (a) A visualization of the *perceptual* distance between all pairs of scene categories in our dataset. The elements in the diagonal represent the perceptual distance *within* the category while all the other elements represent the perceptual distance *between* their corresponding categories. (b) A visualization of the SUN's "good workers" classification confusions between all pairs of scene categories in our dataset. In both cases, scene categories are organized to Natural (top left), Manmade Outdoor (center) and Manmade Indoor (bottom right), separated by black dashed lines. To avoid clutter only a subset of the scene category labels are presented.

are strongly inconsistent with their semantic counterparts. For example, as illustrated in Fig. 2, the "Baseball Field" category is perceptually more related to natural scene categories (e.g., "Desert-Sand", "Beach" and "Field Cultivated") than to most of the manmade categories (e.g., "Castle","Doorway Outdoor" and "Pagoda"), although semantically the opposite holds [18]. Similarly, the "Harbor" category is perceptually more related to several natural scene categories (e.g., "Lake", "Islet") than to many manmade scene categories (e.g., "Street", "Corridor") while semantically the opposite holds again [18] (see Fig. 2). It is this *new* information on scene categories that we wish to exploit for better scene understanding representation and operations.

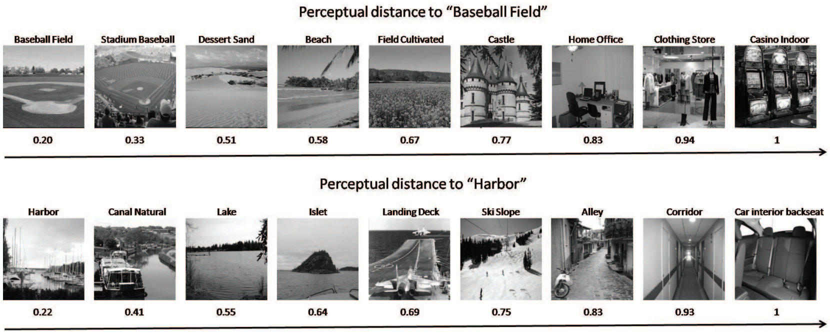


Fig. 2. Our perceptual distance metric for two scene categories examples "Baseball Field" and "Harbor". The other scene categories are labeled with their perceptual distance to the two examples.

2.3 Embedding Categories in Perceptual Space

Our next step is to embed the scene categories along with their perceptual relationships into a possibly lower dimensional perceptual space $\mathbf{R}^c$ such that their Euclidean distances are "perceptually meaningful". One way to carry such embedding 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 [23]. In our case these "distances" are the perceptual distances obtained from human vision and although the dimension c of $\mathbf{R}^c$ can be lower, in our analysis we select $c = 58$ in order to preserve the perceptual distances as much as possible. This choice was mandated by the projection of the PD matrix onto the cone of positive semi-definite matrices by forcing negative eigenvalues to zero.

Let $P = [p_1, ..., p_c] \in \mathbf{R}^{c \times c}$ be a matrix whose columns consist of sought-after scene category prototypes, where p_α is the prototype that represents scene category α. We aim to embed the category prototypes such that the distance $||p_\alpha - p_\beta||_2$ reflects the perceptual distance specified in $PD(\alpha, \beta)$ (i.e., the perceptual distance between categories α and β). More formally, our problem becomes

$$P_{SceNet} = \arg\min P \sum_{\alpha,\beta=1}^{c} (||p_\alpha - p_\beta||^2 - PD(\alpha, \beta))^2 \qquad (1)$$

and it can be solved with metric multi-dimensional scaling [23]. Fig. 3 illustrates the embedding of all the scene categories in our dataset into the first two dimensions of the perceptual space. Interestingly, even with just two dimensions visualized, the results reveal that perceptual relationships do not necessarily conform to their semantic counterparts (e.g., see "Baseball Field", "Gulf Course","Green House Indoor","Phone Booth","Market Outdoor", "Shop Front"). As we demonstrate later (see section 4), the use of this perceptual ontology and space provides significant improvements in scene recognition over the SUN semantic ontology [1], suggesting that the use of the perceptual space over the semantic one should be prioritized in general. At the same time, in agreement with the Spatial Envelope model [24], the first (and most dominant) perceptual dimension appears to be related to the *Naturalness* and *Openness* attributes of the scene. While this can be observed intuitively from the visualization of the obtained perceptual space (see Fig. 3), these findings invite further analysis of the discriminative power of visual attributes (see Sec. 5.1).

Indeed, what are the benefits that such a large-scale perceptual organization may provide over previous perceptual studies that were at much smaller scale with only 8 categories [24, 21, 25]? We argue (and later demonstrate in Sec. 4) that the perceptual space just described is already highly useful in facilitating significant improvements in large-scale scene recognition applications over state-of-the-art methods and existing semantic ontologies. At the same time, since its larger scale set of categories provides much better representation for richness of the real world, the perceptual structure offered by SceneNet could also provide important insights into human scene organization and representation.

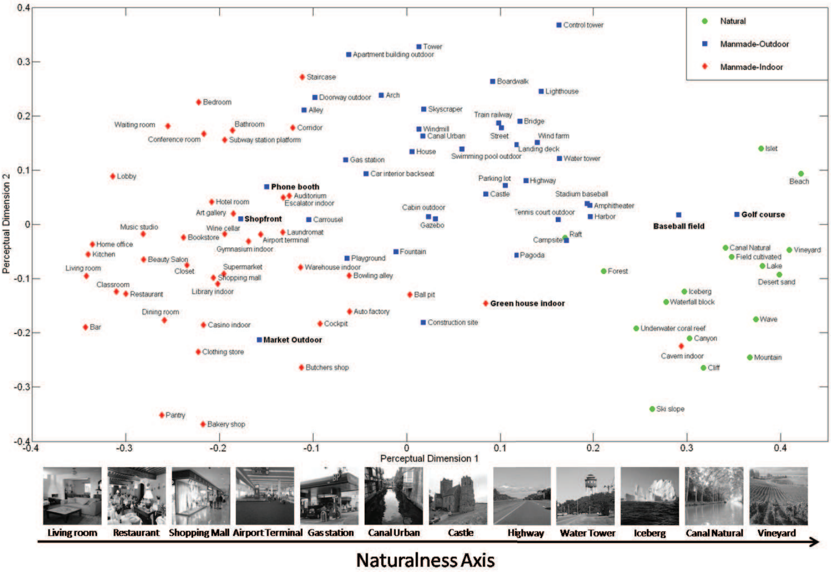


Fig. 3. Visualization of the first two dimensions of the perceptual space. Note that Natural, Manmade outdoor, and Manmade indoor scene categories are colored green, blue, red, respectively. Several categories that are referenced in the text are shown in bold face for faster localization.

In what follows we demonstrate this by exploiting our new perceptual ontology for novel findings about visual attributes and in particular about their discriminative power. For that, we combine SceneNet with the *SUN attribute database* which was recently proposed by Patterson and Hays [26].

3 Large-Scale Scene Recognition with SceneNet

With the SceneNet Database established via experimental analysis as above, we turn to discuss how it may be exploited for large-scale scene recognition. To do so we extend the document taxonomy embedding by Weinberger and Chapelle [20] to allow non-linear ontology embedding via kernels.

3.1 Scene Mapping with Regression

Let a scene image represented as feature vector $x_i \in \mathbf{R}^d$. Once we found a suitable embedding P of the scene category prototypes into $\mathbf{R}^c$, out next step is to find an appropriate linear mapping $W \in \mathbf{R}^{c \times d}$ that maps each input image x_i with category label y_i as close as possible to its category prototype p_{y_i} in the perceptual space. We can find such a linear transformation $z_i = Wx_i$ by setting

$$W = \arg\min W \sum_{i=1}^{n} ||p_{y_i} - Wx_i||^2 + \lambda||W||^2 \tag{2}$$

where n is the number of input images and λ determines the depth of regularization on W, which is necessary to prevent potential overfitting due to the high number of features. The minimization in Eq. 2 is an instance of *linear ridge regression* whose closed-form solution is

$$W = PJX^T(XX^T + \lambda I_d)^{-1} \tag{3}$$

where $I_d \in \mathbf{R}^{d\times d}$ is the identity matrix, $X = [x_1, ..., x_n]$, and $J \in \{0,1\}^{c\times n}$, with $J_{\alpha i} = 1$ if and only if $y_i = \alpha$.

The above formulation can be easily extended to *kernel ridge regression* [27] to use kernels in the following way

$$W = PJ\kappa(x)(K + \lambda I_n)^{-1} \tag{4}$$

where $K \in \mathbf{R}^{n\times n}$ with elements $K_{ij} = \phi(x_i)^T\phi(x_j)$, $\kappa(x) \in \mathbf{R}^n$ with elements $\kappa_i = \phi(x_i)^T\phi(x)$, and ϕ is the feature mapping function.

In order to categorize a new input x_k, we first map it into the perceptual space

$$z_k = Wx_k = PJ\kappa(x_k)(K + \lambda I_n)^{-1}\ . \tag{5}$$

Then, we estimate its label $\hat{y_k}$ as the category with the closet prototype p_α, i.e., via direct nearest neighbor

$$\hat{y_k} = \arg\min \alpha ||p_\alpha - z_k||^2 \tag{6}$$

3.2 Large Margin Scene Mapping

So far we have learned the scene category prototypes P based on the SceneNet-100 ontology (i.e., directly from human vision and independent of the input data X) and found a mapping W that maps each input scene closest to the prototype of its category in the perceptual space. Still, a better and more robust generalization would allow for the correct prototype p_i to lie much closer to z_i than any other prototype p_α *by a large margin*. Moreover, it would be also preferable if perceptually dissimilar prototypes would be further separated by a *larger* margin than those that are more perceptually related. In the following we briefly describe the large margin formulation [20] to learn P and W *jointly* for better generalization.

Following Eq. 4, let us define the following matrix A:

$$A = J\kappa(x)(K + \lambda I_n)^{-1}\ . \tag{7}$$

Eq. 4 and 7 suggest that $W = PA$ and that A is completely independent of P and can be computed directly from the input data X. Scene category prototype p_α and query z_i can now be rephrased as follows:

$$p_\alpha = Pe_\alpha \quad \mathit{and} \quad z_i = Px'_i \tag{8}$$

where $x'_i = Ax_i$ and $e_\alpha = [0, ..., 1, ..., 0]^T$ (i.e., vector with all zeros and a single 1 in the α^{th} position). This allow us to reduce the problem to a single optimization problem to determine P while enforcing large margin constraints with respect to the perceptual relationships between scene categories (i.e., $PD_{y_i\alpha}$) and using regularization with weight $\mu \in [0,1]$ to ensure that P will be as similar as possible to P_{SceNet} (cf. Eq. 1). We hence define the following constrained optimization

$$\begin{aligned} &\arg\min P (1-\mu)\sum_{i,\alpha} \xi_{i\alpha} + \mu ||P - P_{SceNet}||^2 \quad \textbf{subject to} \\ &(1) \quad ||P(e_{y_i} - x'_i)||^2 + PD_{y_i\alpha} \leq ||P(e_\alpha - x'_i)||^2 + \xi_{i\alpha} \\ &(2) \quad \xi_{i\alpha} \geq 0 \end{aligned} \tag{9}$$

where $PD_{y_i\alpha}$ now represents the (large) margin we wish to enforce on the correct classification while the slack variables $\xi_{i\alpha}$ absorb the amount of violation of prototype $p_{\alpha\neq y_i}$ into the margin of the correct prototype p_{y_i} [20].

As is later demonstrated in Sec. 4, the use of regularization term in the objective function is necessary to prevent overfitting due to the high number of features, since while the training input data might differ from the test data, the perceptual ontology remains the same. While the constraints in Eq. 9 are quadratic with respect to P and the optimization is therefore not convex, we can make Eq. 9 convex by defining $Q = P^T P$ and rewriting all distances in terms of Q while requiring that Q is positive semi-definite. With

$$||P(e_\alpha - x'_i)||^2 = (e_\alpha - x'_i)^T Q (e_\alpha - x'_i) = ||e_\alpha - x'_i||^2_Q \tag{10}$$

we therefore rewrite the final convex optimization problem as follows:

$$\begin{aligned} &\arg\min Q (1-\mu)\sum_{i,\alpha} \xi_{i\alpha} + \mu ||Q - Q_{SceNet}||^2 \quad \textbf{subject to} \\ &(1) \quad ||(e_{yi} - x'_i||^2_Q + PD_{y_i\alpha} \leq ||e_\alpha - x'_i||^2_Q + \xi_{i\alpha} \\ &(2) \quad \epsilon_{i\alpha} \geq 0 \\ &(3) \quad Q \geq 0 \end{aligned} \tag{11}$$

where $Q_{SceNet} = P^T_{SceNet} P_{SceNet}$. This optimization is a particular instance of semi-definite program (SDP) [28] that can be solved very efficiently with special purpose sub-gradient solvers [29]. Once the optimal solution Q^* is found, one can obtain P with svd $Q^* = P^T P$ and the mapping W from $W = PA$.

4 Large-Scale Scene Categorization

While the goal of this paper and the SceneNet ontology and database is not necessarily limited to improved scene categorization, in this section we demonstrate how the use of the SceneNet-100 ontology embedding facilitates significant improvements in this central and popular task. To do so, we compared our approach, abbreviated here as *SceNet-Ontem*, to the one-vs-all Support Vector Machines (*SVM 1/all*) using publicly available code [1] with the descriptor

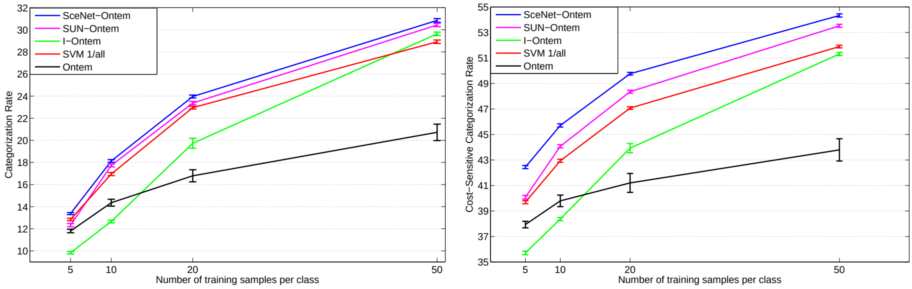


Fig. 4. Scene Categorization: Performance of all discussed algorithms (*SceNet-Ontem,SUN-Ontem,SVM 1/all*, *I-Ontem*, and *Ontem*) as the number of training examples is increased. **Left**: The standard categorization rate that treats each misclassification equally. **Right**: The cost sensitive categorization rate that treats each misclassification according to the perceptual distances between scene categories. Error bars represent standard error of the means.

and kernel defined as above. Additionally, we show that this improvement results from the very specific ontology represented by SceneNet-100 which was inferred experimentally and reflects human perception. To do so, we also compared *SceNet-Ontem* to *I-Ontem*, an instance of our ontology embedding where the SceneNet ontology is *ignored* and P is set to be the identity matrix $I \in \mathbb{R}^{c \times c}$ such that all category prototypes are placed in constant distance from each other in the perceptual space. Furthermore, we also compared *SceNet-Ontem* to *SUN-Ontem*, an instance of our ontology embedding where the manually defined sematic ontology from SUN is used [1]. Finally, we tested another control classifier, dubbed here as *Ontem*, where the regularization term in the SDP (which is used to enforce similarity between P and P_{SceNet}) is completely ignored by setting $\mu = 0$.

In all cases we randomly split each category to disjoint training and testing sets, with $n_{training} = 5, 10, 20, 50$ and $n_{test} = 50$. The same sets where then used with the five algorithms (i.e., *SceNet-Ontem,SUN-Ontem ,SVM 1/all, I-Ontem,Ontem*) and repeated 20 times (to control for the random selection of samples). The GIST descriptor that was proposed specifically for scene recognition tasks [24] was used with an RBF kernel using the code available online [1]. We set the regularization weights to $\lambda = 1$ for the kernel ridge regression and $\mu = 0.9$ for the SDP. Preliminary experiments have shown that regularization was important but the exact settings of the λ and μ had no crucial impact, except for the need for μ to be closer to one than to zero to insure that P will be similar enough to P_{SceNet}. We evaluated the performance of two measures of categorization accuracy (each measure treats the misclassification differently): (1) the *conventional* categorization rate that weighs each misclassification equally; (2) the *cost sensitive* categorization rate that weighs each misclassification according to the perceptual distance between scene categories. The latter measure is inspired by the observation that quite often the implications of confusing certain classes is less critical than others. For example, it is easy to conceive an application where it is significantly worse to misclassify a *coastal* scene as a *kitchen* rather than a *lake*.

A comparison of the five algorithms and two measures of performance is provided in Fig. 4. As the results show, the use of the SceneNet ontology embedding yields significant improvement over *SVM 1/all* in all training set sizes. The graphs also show that the ontology used cannot be arbitrary but rather it must reflect the true relations between scene categories. Indeed, when all scene categories have constant distance from each other (as in *I-Ontem*), or when P is not required to be similar to P_{SceNet} (as in *Ontem*), performance drops significantly. Finally, while using semantic ontology (cf. *SUN-Ontem*) may improve performance compared to *SVM 1/all*, the use of the SceneNet ontology yields significantly better performance in all training set sizes.

5 Perceptual insights

Apart from significantly better computational results, the SceneNet database could also provide important findings in human scene organization and representation. In what follows we demonstrate this by exploiting our new perceptual ontology for novel findings about the discriminative power visual attributes and the typicality of scenes.

5.1 Discriminative power of visual attributes

In their recent attempt to enable deeper understanding of scenes, Patterson and Hays [26] proposed the SUN attribute database that spans over 700 categories and 14,000 images with 102 discriminative attributes related to materials, surface properties, lighting, functions/affordance, and spatial envelope properties. While they reported that scene category can be predicted only from scene attributes, using SceneNet we now attempt to determine which among these attributes are the most discriminative, or more generally, to obtain insights about the discriminative power of all attributes. Specifically, we argue that the more discriminative attributes account for most of the distance between scene categories in our perceptual space while less discriminative attributes have only minor effect on the perceptual distance between scene categories. In other words, exploring the interaction between these two databases may reveal this new information very explicitly.

Following the information in the SUN attribute database, let a scene image be represented as attribute feature vector $a \in \mathbf{R}^{102}$. For each pair of scenes a_i and a_j from two distinct scene categories α and β from SceneNet-100, we calculate the vector $d_{ij} \in \mathbf{R}^{102}$ of their absolute pairwise differences. Since d_{ij} reflects the attributes that distinguish scenes a_i and a_j, we refer to it as the attribute-distance vector between scenes a_i and a_j. Next, we trained a support vector regression (ϵ SVR) to map each attribute-distance vector d_{ij} to the perceptual distance value between their corresponding categories α and β (i.e., $PD_{\alpha\beta}$). With the trained support vector regression we could now predict the discriminative power of each visual attribute separately with the input $e_z = [0, ..., 1, ..., 0]^T$ (i.e., vector with all zeros and a single 1 in the z^{th} position for attribute z). Fig. 5 plots these results for all visual attributes in the SUN attribute database, sorted by discriminative power.

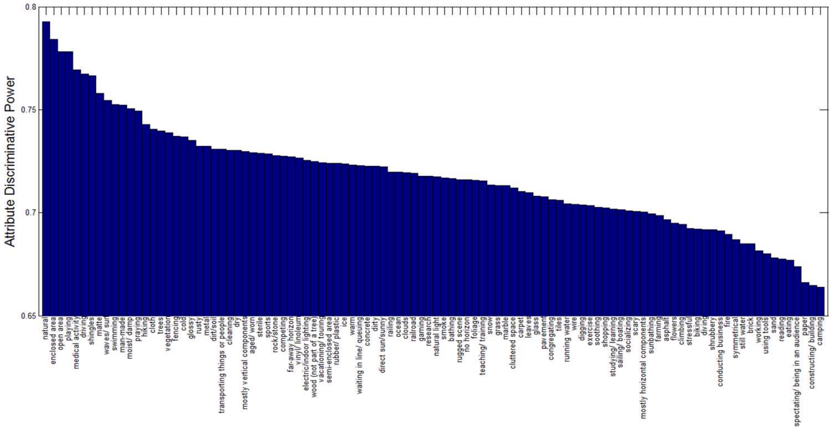


Fig. 5. The discriminative power of the visual attributes in the SUN attribute database. Consistent with the spatial envelope model [24], the most discriminative attributes are “Natural” and ”Openness” (i.e., enclosed area, open area). Here, however, with a number of scene categories and attributes that is an order of magnitude larger than that of [24], we provide a rigorous perceptual basis to support and validate this claim.

Consistent with the spatial envelope model [24], the most discriminative attributes are “Natural” and ”Openness” (i.e., enclosed area, open area). Here, however, with a number of scene categories and attributes that is an order of magnitude larger than [24], we provide a rigorous perceptual basis to support and validate this claim. More significantly, we provide a full evaluation of the discriminative power for the most comprehensive list of visual attributes available to-date, which enables deeper understanding of visual attributes and their relations to human perception, and could possibly facilitate better attribute-based scene representation for scene recognition.

5.2 Typicality of scenes

One of the most robust findings in categorization is that category membership is graded and that humans seem to consistently identify typical and atypical exemplars of a category [30, 31]. More importantly, there is a large body of work supporting the influence of typicality on categorization (see [32] for a detailed review). For example, it has been found that observers response time is faster for queries involving typical exemplars (e.g., “is a robin a bird”) than for atypical exemplars (e.g., “is a chicken a bird”) [33], and that learning of category representations is faster when typical rather than atypical exemplars are presented earlier in the sequence [34]. Arguing that a proper scene representation should take these findings into account and be consistent with them, we use our perceptual space scene representation to obtain a new perceptual typicality measure that correlates highly with the typicality ranking of humans.

The perceptual space scene representation (as opposed to discriminative methods such as SVM) has the advantage of representing a soft decision about the degree to which an image belongs to a category. We measure the image typicality by computing the distance between scene images and their categorical prototypes in perceptual space. Examples of the most typical and atypical images by our approach are shown in Fig 6.

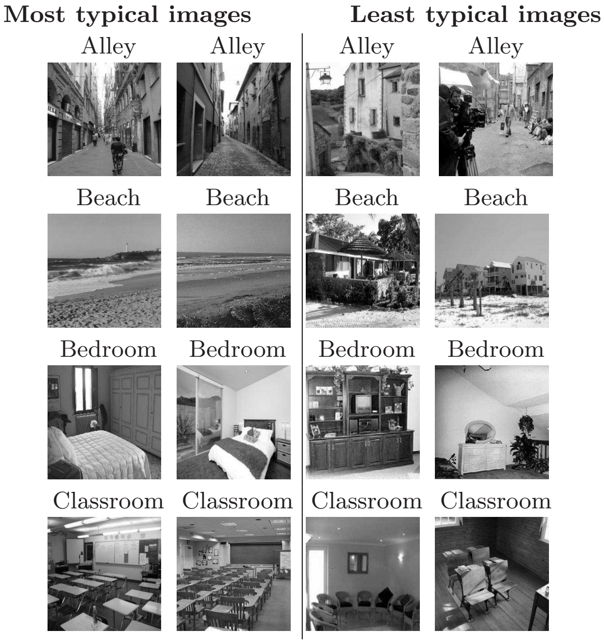
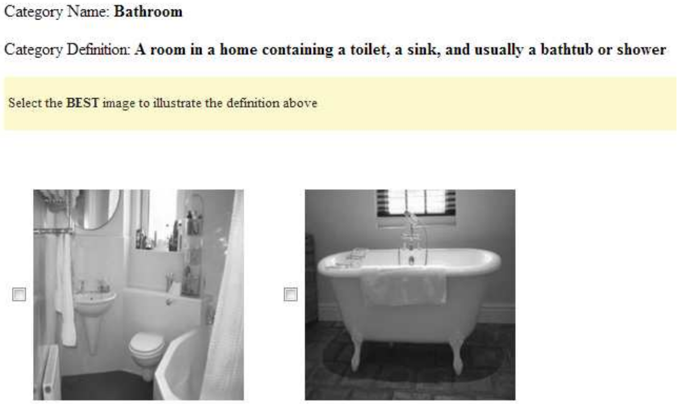


Fig. 6. Examples of the most typical and atypical images by our approach.

Next, we conducted a psychophysical experiment to compare the typicality measure based on the SceneNet perceptual space scene representation (SceneNet typicality measure) with the typicality ranking of humans. In particular, we presented workers on Amazon Mechanical Turk (AMT) with the *Image Typicality Task*, where participants were given the name of a scene category from the SUN-100 database, a short definition of the scene category, and two images. Workers were asked to select which of the two images best described the name and definition (one of the two images was drawn from the 10 most typical images by our approach and the other was drawn from the 10 most atypical images by our approach). A total of 42 workers from AMT (with better than 97% approval rate, at least 5000 approved HITs, and located in the United States) performed the task to provide a large pool of 1000 trials. Workers were compensated with $0.02 per HIT. An sample trial from the Image Typicality Task is shown in Fig 7.



Category Name: **Bathroom**

Category Definition: **A room in a home containing a toilet, a sink, and usually a bathtub or shower**

Select the **BEST** image to illustrate the definition above

□ □

Fig. 7. An example of a trial in the Image Typicality Task. In each trial, participants were given the name of a scene category from the SUN-100 database, a short definition of the scene category, and two images. Workers were asked to select which of the two images best described the name and definition (one of the two images was drawn from the 10 most typical images by our approach and the other one was drawn from the 10 most atypical images by our approach)

Having all worker responses in the Image Typicality Task, we then assessed the degree of agreement between the SceneNet typicality measure and the human subjects' typicality ranking. Our analysis revealed that scenes that humans rate as more typical examples of their category are more likely to be close to their categorical prototype in the perceptual space. Indeed, participants selected an image from the most typical scene images by our approach in 84.38% of the trials, indicating that the SceneNet perceptual space scene representation and the SceneNet typicality measure are perceptually plausible.

6 Conclusion

In this paper we argue that in order to advance the field of scene understanding a proper knowledge of the *perceptual* ontology of scene categories is required. We have proposed such an ontology and provided SceneNet-100, an ontological database of 100 scene categories that was derived directly from human vision through a large-scale human study. The SceneNet ontology and database organizes scene categories according to their *perceptual* relationships and provides a lower dimensional scene representation with "perceptually meaningful" Euclidean distances. We show that the use of SceneNet facilitates significant improvements in large-scale scene categorization and provides important insights into human scene representation and organization for the benefit of future exploration of scene understanding.

References

1. Xiao, J., Hays, J., Ehinger, K., Oliva, A., Torralba, A.: Sun database: Large scale scene recognition from abbey to zoo. In: CVPR. (2010)
2. : SceneNet:An Online Perceptual Ontology Database for Scene Understanding. (2013) Anonymous URL. Concealed for blind review.
3. Fei-Fei, L., Perona, P.: A bayesian hierarchy model for learning natural scene categories. In: CVPR. (2005)
4. Bosch, A., Zisserman, A., Munoz, X.: Scene classification via pLSA. In: ECCV. (2006) 517–530
5. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: CVPR. (2006) 2169–2178
6. Griffin, G., Perona, P.: Learning and using taxonomies for fast visual categorization. In: CVPR. (2008)
7. Bart, E., Porteous, I., Perona, P., Welling, M.: Unsupervised learning of visual taxonomies. In: CVPR. (2008)
8. Ahuja, N., Todorovic, S.: Learning the taxonomy and models of categories present in arbitrary images. In: ICCV. (2007)
9. Schmid, C.: Constructing category hierarchies for visual recognition. In: ECCV. (2008)
10. Sivic, J., Russell, B., Zisserman, A., Freeman, W., Efros, A.: Unsupervised discovery of visual object class hierarchies. In: CVPR. (2008)
11. Li, L., Wang, C., Lim, Y., Blei, D., Fei-Fei, L.: Building and using a semantivisual image hierarchy. In: CVPR. (2010)
12. Marszalek, M., Schmid, C.: Semantic hierarchies for visual object recognition. In: CVPR. (2007)
13. Torralba, A., Fergus, R., W.T, F.: 80 million tiny images: a large dataset for nonparametric object and scene recognition. IEEE Trans. Pattern Anal. Mach. Intell. **30**(11) (2008) 1958–1970
14. Fergus, R., Bernal, H., Weiss, Y., Torralba, A.: Semantic label sharing for learning with many categories. In: ECCV. (2010)
15. Deselaers, T., Ferrari, V.: Visual and semantic similarity in imagenet. In: CVPR. (2011) 1777–1784
16. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR. (2009)
17. Verma, N., Mahajan, D., Sellamanickam, S., Nair, V.: Learning hierarchical similarity metrics. In: CVPR. (2012)
18. Miller, G.: Wordnet: A lexical database for english. In: Communications of the ACM. (1995)
19. Deng, J., Berg, A., Fei-Fei, L.: Hierarchical semantic indexing for large scale image retrieval. In: CVPR. (2011)
20. Weinberger, K., Chapelle, O.: Large margin taxonomy embedding for document categorization. In: NIPS. (2008) 1737–1744
21. Kadar, I., Ben-Shahar, O.: Small sample scene categorization from perceptual relations. In: CVPR. (2012) 2711–2718
22. Rousselet, G.A., Fabre-Thorpe, M., Thorpe, S.J.: Parallel processing in high-level categorization of natural images. Nature Neuroscience **5**(7) (2002) 629–630
23. Torgerson, W.S.: Multidimensional scaling: theory and method. Psychometrika **17**(6) (1952) 401–419

24. Oliva, A., Torralba, A.: Modeling the shape of the scene: A holistic representation of the spatial envelope. Int. J. Comput. Vision **42**(3) (2001) 145–175
25. Greene, M., Oliva, A.: Forest before the trees: the precedence of global features in visual perception. Cognit. Sci. **58** (2009) 137–179
26. Patterson, G., Hays, J.: SUN attribute database: Discovering, annotating, and recognizing scene attributes. In: CVPR. (2012)
27. Saunders, C., Gammerman, A., Vovk, V.: Ridge regression learning algorithm in dual variables. In: ICML. (1998) 515521
28. Boyd, S., Vandenberghe, L., eds.: Convex Optimization. Cambridge University Press (2004)
29. Weinberger, K., Saul, L.: Fast solvers and efficient implementations for distance metric learning. In: ICML. (2008) 1160–1167
30. Vogel, J., Schiele, B.: Semantic typicality measure for natural scene categorization. In: Annual Pattern Recognition Symposium. (2004)
31. Ehinger, K., Xiao, J., Torralba, A., Oliva, A.: Estimating scene typicality from human ratings and image features. In: Proceedings of the 33rd Annual Conference of the Cognitive Science Society. (2011) 2562–2567
32. Murphy, G.L., ed.: The big book of concepts. MIT Press (2002)
33. Rosch, E.: Cognitive representations of semantic categories. J. Exp. Psych. (1975)
34. Mervis, C., Pani, J.: Acquisition of basic object categories. Cognit. Sci. **12** (1980)