

Visual and Semantic Similarity in ImageNet

Thomas Deselaers^{1,2} and Vittorio Ferrari¹

¹Computer Vision Laboratory
ETH Zurich, Switzerland
{deselaers, ferrari}@vision.ee.ethz.ch

²Google
Zurich, Switzerland
deselaers@google.com

Abstract

Many computer vision approaches take for granted positive answers to questions such as “Are semantic categories visually separable?” and “Is visual similarity correlated to semantic similarity?”. In this paper, we study experimentally whether these assumptions hold and show parallels to questions investigated in cognitive science about the human visual system. The insights gained from our analysis enable building a novel distance function between images assessing whether they are from the same basic-level category. This function goes beyond direct visual distance as it also exploits semantic similarity measured through ImageNet. We demonstrate experimentally that it outperforms purely visual distances.

1. Introduction

Categories are a central subject in both computer vision and cognitive science. Cognitive psychology [36] studies categories as *semantic* units in the human mind, and investigates questions such as “How do humans define categories?” [36], “How are categories represented (in the human mind)?” [36], and “Are there conceptual prototypes for a category?” [30, 31]. The ability to reason at the category level is even considered a basis of human intelligence [20].

For the human visual system, cognitive science has found positive answers to questions such as “Do semantic categories form clusters in visual space?”, “Are there visual prototypes for a semantic category?”, “Is visual similarity correlated to semantic similarity?”, and “Are semantic categories visually separable?”. However, the answers to these questions are currently unclear when visual similarity is measured by modern computer vision techniques. In spite of this, many recognition systems implicitly build on the assumption of positive answers. In this paper we study experimentally whether these assumptions hold. We investigate the relations between visual and semantic category similarity on the recent ImageNet dataset [7]. This large-scale dataset contains about 10 Million images in about 15’000 categories organized according to the semantic hi-

erarchy of WordNet [11] (fig. 1). More precisely, we study the following aspects:

(i) We analyze how the visual variability within a category changes with depth in the hierarchy, i.e. the size of its semantic domain. In particular we test whether a smaller semantic domain [15] corresponds to a smaller visual variability (sec. 3).

(ii) We determine a visual prototype for every category and measure how well it represents the category as a whole (sec. 3). This analysis ties in with prototype theory [30, 31] from cognitive science. It states that for sufficiently specific categories, e.g. *bird*, humans agree on a single prototype defined by a typical shape and attributes such as *can fly* and *feathered*. For broader categories instead, such as *animal*, this is not the case.

(iii) We measure the relation between semantic and visual similarity (sec. 4). In cognitive psychology, categories are typically defined by grouping “similar objects”, and super-categories by grouping “similar categories” [36]. Are these conceptual similarities in categories defined by humans reflected in the visual similarity between images of these categories? E.g. are images of different dogs more similar than images of dogs and cows, and in turn more similar than images of cows and motorbikes? We attempt to answer the question whether semantic similarity implies visual similarity, which is assumed by most visual recognition approaches and which has been shown to be true for the human visual system [30, 31].

(iv) We analyze how within-class and between-class visual similarities change as a function of how broadly classes are semantically defined (sec. 4). Our analysis focuses on how well such classes are visually separable. It provides evidence for answering whether computer vision algorithms have a chance to classify across semantically meaningful class boundaries. While humans can distinguish tens of thousands of categories in visual as well as in semantic space [3], it is currently not clear whether it is possible to scale computer vision algorithms to that extent and if current image descriptors are powerful enough.

The insights gained during the above investigations en-

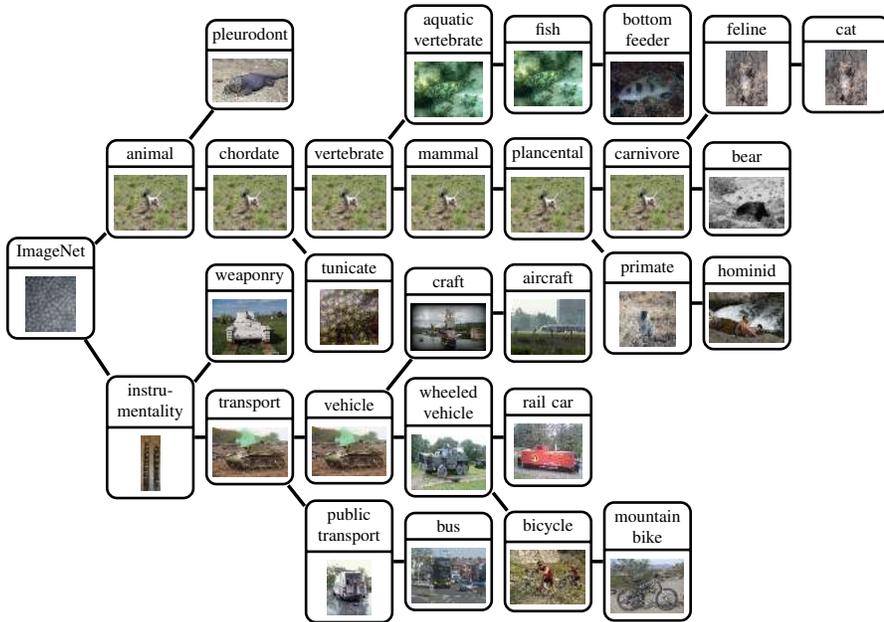


Figure 1: **The ImageNet hierarchy.** Some paths in the hierarchy with their representative images determined as in sec. 3.

able to build a new distance between pairs of images which assesses whether they show the same basic-level category (e.g. "car", "dog"; these are important as they are the level at which human most frequently reason [32]). As opposed to previous works on distance functions measuring purely the visual similarity between the two query images [1, 13, 14, 22, 24, 28], our distance employs ImageNet as a large pool of background data enabling to make additional semantic connections beyond direct visual similarity (sec. 5). As we experimentally demonstrate, this new distance function outperforms pure visual distances. This makes it valuable for object recognition, image annotation, and image retrieval, where it can be used in a nearest neighbor classifier or as a kernel in an SVM.

The paper is structured as follows. After introducing the ImageNet dataset (sec. 2), we analyze the visual scale of categories as a function of their semantic domain (sec. 3) and the relation between visual and semantic similarity (sec. 4). Section 5 presents our novel distance between two images and evaluate it experimentally on ImageNet. Related work is discussed in sec. 6 and concluding remarks are given in sec. 7.

2. The ImageNet Dataset

We build our analysis on the ImageNet dataset [7] (Fall 2009 release). ImageNet contains 9'353'897 images in 14'791 categories organized according to the semantic hierarchy of WordNet [11]. A category in ImageNet corresponds to a *synonym set* (synset) in WordNet. ImageNet covers a subset of the nouns of WordNet, organized in 12

top-level categories, e.g. *animal*, *instrumentality* (fig. 1). Additionally, for 141'731 images from 548 synsets bounding boxes are available¹. Compared to other large datasets, e.g. TinyImages [37], ImageNet offers two advantages: (i) the images are in higher resolution. (ii) after downloading the images from image search engines, they were manually verified to contain the relevant concepts using Amazon Mechanical Turk (AMT) [7]. Every node (category) in the hierarchy contains on average 632 images unique to that node. Moreover, every node also contains all images in its subtrees (subcategories). In the example in fig. 1, the "animal" node includes all images of its children, e.g. "chordate", "vertebrate", "mammal", etc., plus some images of its own.

3. The Visual Scale of Categories

First we investigate the *visual scale* of categories at different depths in the hierarchy. This measures how much visual variability there is among instances of a category. We represent images using GIST [25], which was shown to describe whole images well [8, 18]. GIST consists of Gabor orientation histograms computed over cells in a regular grid. (fig. 3f visualizes a GIST descriptor of the image in the 'animal' node of fig. 1). We use GIST with the default parameters [25] (3 color planes, 4x4 cells, and 3 scales with 8, 8 and 4 orientations respectively, giving 960 dimensions).

We measure the visual scale r_S of a category S as the average distance between its mean GIST descriptor μ_S and

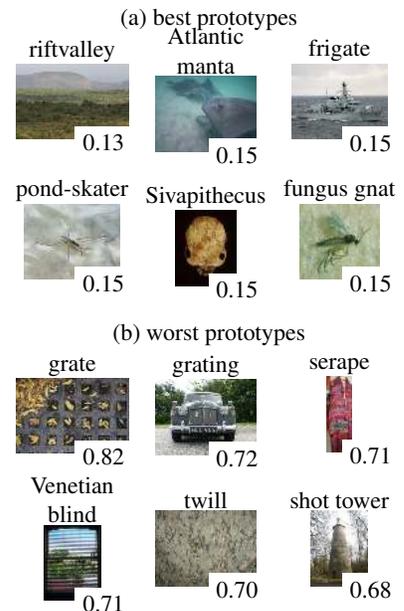


Figure 2: **Example prototype images.** Best (a) and worst (b) prototypes. Category names are above the images, q -scores in the bottom right corner.

¹These bounding boxes come from the ImageNet Spring 2010 release.

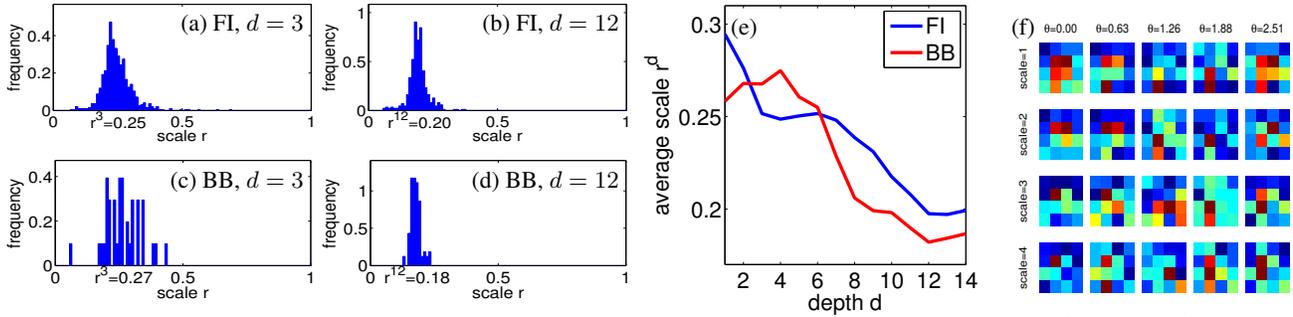


Figure 3: **Scales of categories at different depths.** (a,b): histograms of visual scale at depth $d = 3, 12$ with $r^3 = 0.25, r^{12} = 0.20$ for FI; (c,d): histograms of visual scale at depth $d = 3, 12$ with $r^3 = 0.27, r^{12} = 0.18$ for BB (these histograms are sparser since only 142K of the 10M images are annotated with BB) (e): average visual scale r^d as a function of depth d for FI and BB. (f): GIST descriptor with 5 orientations in 4 scales for the image in the node “animal” of fig. 1. Every subpanel shows the average Gabor responses on a 4×4 grid on the image.

all images in \mathcal{S}

$$r_{\mathcal{S}} = \frac{1}{|\mathcal{S}|} \sum_{I \in \mathcal{S}} D(I, \mu_{\mathcal{S}}) \quad (1)$$

where $D(I, \mu_{\mathcal{S}})$ is the squared Euclidean distance.

We study the distribution of visual scales among all categories at a given depth in the hierarchy for both full images (FI) and objects cropped according to their bounding-box annotation (BB). We measure how $r_{\mathcal{S}}$ changes with depth both when measured on FI and on BB.

As fig. 3a-d show for two depths, the scale of categories at a depth is roughly Gaussian and very peaked, for both FI and BB. Moreover, the average visual scale r^d over the categories at depth d steadily decreases with d (fig. 3e). Interestingly, this corresponds to categories covering smaller and smaller semantic domains (e.g. at depth 3 there are *animal/chordate* and *instrumentality/transport*; while at depth 11 there are *animal/.../Chihuahua* and *instrumentality/.../minivan*). These results show that categories with smaller semantic domains also have smaller visual variability. This confirms human intuition and provides experimental support for this basic assumption made in computer vision. We performed a similar analysis with other definitions of visual scale and observed the same behavior (e.g. the average dimension-wise variance of the descriptors)

Closely related to the visual scale of a category is how well it can be represented using a single prototype image. We select as the prototype $\hat{\mu}_{\mathcal{S}}$ of a category \mathcal{S} the image I minimizing the sum of squared distances (SSD) to all images in \mathcal{S} which can be computed efficiently as the image closest to the synset mean $\mu_{\mathcal{S}}$

$$\hat{\mu}_{\mathcal{S}} = \arg \min_{I' \in \mathcal{S}} \sum_{I \in \mathcal{S}} D(I, I') = \arg \min_{I \in \mathcal{S}} D(I, \mu_{\mathcal{S}}) \quad (2)$$

Further, we consider $q(\hat{\mu}_{\mathcal{S}}) = \frac{1}{|\mathcal{S}|} \sum_{I \in \mathcal{S}} D(I, \hat{\mu}_{\mathcal{S}})$ as a measure of quality of $\hat{\mu}_{\mathcal{S}}$ (i.e. normalized SSD to all images in \mathcal{S}). We did this analysis for FI. To ensure stable

estimations, we consider only categories with at least 100 images. Visually compact categories will be described well using a single prototype. The prototypes for the categories with the best and worst q -scores are shown in fig. 2. The best prototypes come from specific natural categories. For example the images of “rift valley”² are mostly landscapes with sky and grass; the category “Atlantic manta”³ shows underwater images with large dark trapezes in the middle. Interestingly, the categories with the worst prototypes are man-made objects defined by their function, which have large visual variability. For example “grate”⁴ and “grating”⁵ contain various kind of grates applied over very different objects; the category “serape”⁶ contains clothing in different colors, some alone, some worn by persons in various locations.

Further prototypes are shown in fig. 1. Interestingly, the prototype for the entire ImageNet is a regular isotropic gray texture. Although it is a good average image, it has little semantic meaning. For the basic-level categories (e.g. “cat”, “bus”) and for some of the broader categories (e.g. “wheeled vehicle”) the prototypes represent their respective categories well. We released the prototypes online⁷.

4. Relationship between Semantic and Visual Similarity

We investigate how semantic distances between categories defined on the WordNet hierarchy relate to visual distances in ImageNet. For our analysis we choose the Jiang and Conrath semantic distance (JC) [19], which was shown to outperform other semantic distances on WordNet for several natural language processing tasks [4]. The JC distance between two categories \mathcal{S} and \mathcal{T} in the hierarchy is defined

²<http://www.image-net.org/synset?wnid=n09410224>

³<http://www.image-net.org/synset?wnid=n01500476>

⁴<http://www.image-net.org/synset?wnid=n03454536>

⁵<http://www.image-net.org/synset?wnid=n03454707>

⁶<http://www.image-net.org/synset?wnid=n04173907>

⁷<http://www.vision.ee.ethz.ch/calvin/imagenet>

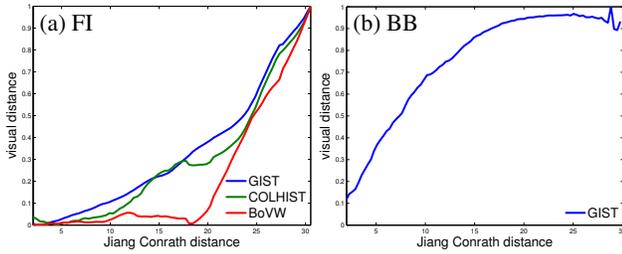


Figure 4: **Relationship between semantic and visual distance** for (a) FI using 3 visual descriptors (b) BB using GIST. Each line shows the average visual distance in 100 semantic distance bins.

as

$$D^{JC}(\mathcal{S}, \mathcal{T}) = 2 \log(p(\text{Iso}(\mathcal{S}, \mathcal{T}))) - (\log(p(\mathcal{S})) + \log(p(\mathcal{T}))) \quad (3)$$

where $p(\mathcal{S})$ is the percentage of all images in \mathcal{S} , $\text{Iso}(\mathcal{S}, \mathcal{T})$ is the lowest superordinate, i.e. the most specific common ancestor of \mathcal{S} and \mathcal{T} . For example in fig. 1, $\text{Iso}(\text{fish}, \text{carnivore}) = \text{vertebrate}$, $D^{JC}(\text{carnivore}, \text{fish}) = 9.04$, and $D^{JC}(\text{carnivore}, \text{primate}) = 3.87$.

We measure the visual distance $D^V(\mathcal{S}, \mathcal{T})$ between two categories \mathcal{S}, \mathcal{T} as the average distance between the mean descriptor $\mu_{\mathcal{S}}$ of \mathcal{S} and all images in \mathcal{T}

$$D^V(\mathcal{S}, \mathcal{T}) = \frac{1}{|\mathcal{T}|} \sum_{I \in \mathcal{T}} D(\mu_{\mathcal{S}}, I) \quad (4)$$

In our analysis, we do not consider pairs of categories where one is an ancestor of the other, e.g. we consider pairs such as *aquatic vertebrate* and *mammal* but not pairs such as *aquatic vertebrate* and *chordate*. This avoids artificially biasing estimations, as ancestors include all images in their descendants’ categories.

We analyse the relationship between D^{JC} and D^V over all pairs of categories for full images (FI, fig. 4a) and objects cropped according to their bounding-box annotation (BB, fig. 4b). Several interesting phenomena can be observed (GIST curves). First, for both FI and BB, visual distance continuously grows with semantic distance. Second, the visual distances grow at different speeds in different ranges of semantic distance because the background is included in FI and not in BB: (i) at low semantic distances ($D^{JC} < 10$), different categories often share similar background which dilutes measurements of visual distances on FI (e.g. dogs and horses). Measurements made on BB instead better reflect the true dissimilarity of the category instances themselves. This is the reason why the BB curve starts out much more rapidly than the FI curve; (ii) at intermediate semantic distances ($10 < D^{JC} < 20$), backgrounds are more varied (e.g. dogs and fish) and this effect weakens, as reflected by the decreasing slope of the BB plot. (iii) at high semantic distances ($D^{JC} > 20$), categories appear in radically different environments (animal vs man-made objects). This

explains the increasing slope on FI. For BB instead, the visual distance converges indicating that for greater semantic distances all categories are equally visually dissimilar.

For completeness, fig. 4a also reports curves for LAB color histograms (COLHIST) and bag-of-visual-words (BoVW) histograms [5] compared with the χ^2 distance. For BoVW, we use SURF descriptors [2] extracted at SURF interest points and quantize them into 2000 visual words with k-means. The curves follow a similar trend to the GIST one.

In conclusion these results demonstrate that visual similarity as measured by computer vision descriptors truly conveys semantic similarity, analog to what shown for human perception [30]. This relationship is particularly strong when measurements are focused on the categories themselves, ignoring backgrounds. This confirms that visual recognition algorithms may benefit from explicitly localizing category instances in the images.

We now analyze how visual distances within a class and between classes change as a function of how classes are semantically defined, i.e. how broad is the *semantic span* of a class (fig. 5). For a given semantic distance x , we consider all pairs of categories (\mathcal{S}, \mathcal{T}) with $D^{JC}(\mathcal{S}, \mathcal{T}) \leq x$ to belong to the same class, and all pairs ($\mathcal{S}', \mathcal{T}'$) with $D^{JC}(\mathcal{S}', \mathcal{T}') > x$ to belong to different classes (we call this the *semantic span*). For example, at $D^{JC} = 5$, *craft* and *wheeled vehicle* are in the same class, while *craft* and *railcar* are not. At $D^{JC} = 10$, *craft* and *railcar* are in the same class, but *craft* and *cat* are not.

Using this definition of a class, we plot the average within-class visual distances (on GIST), between-class visual distances, and the difference between the two as a function of semantic span (for FI in fig. 5a and for BB in fig. 5b). Remarkably, the average within-class visual distance is smaller than the average between-class visual distance for all but the greatest semantic spans. This suggests that visual classification across semantically-defined class boundaries is feasible for all relevant semantic spans (i.e. how semantically broadly classes are defined). This raises hope that computer vision methods will eventually solve most semantic classification tasks. We performed the same analysis using other semantic distances such as Lin and Resnik-Similarity [4] and other visual descriptors and found similar results.

5. Are two images in the same basic-level category?

The question whether two images show an object of the same class is a fundamental problem in computer vision [1, 14, 22, 24]. As classes, typically researchers are interested in basic-level categories [32], such “car” and “dog”, which are most relevant for humans (as opposed to general categories, such as “animal”, and specific ones such as “chinchilla”). Having a general comparison function between

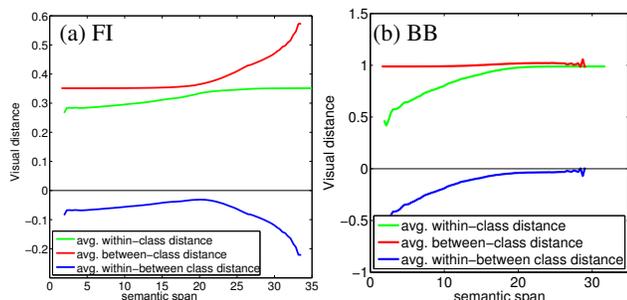


Figure 5: . The within-class distance, the between-class distance, and the difference between them as a function of semantic span for FI (a) and BB (b).

pairs of images that decides whether these show objects of the same basic-level category is useful for image retrieval [8, 13, 38], auto-annotation [17], and object recognition [2, 5, 10, 13, 23, 42], where it can be used in a nearest-neighbor classifier or plugged as a kernel into an SVM.

5.1. ImageNet Distance Between Two Images

We propose here a novel distance function between two images I_i, I_j . Different from previous works [1, 13, 14, 22, 24, 28], it is not based purely on the visual similarity of I_i, I_j , but also exploits semantic similarity as measured through ImageNet. ImageNet acts as a large pool of background data enabling to define a semantic distance between the two images. In a nutshell, the method works as follows. For both input images I_i and I_j , we first search their nearest neighbors \mathcal{N}_i and \mathcal{N}_j in ImageNet using visual distances. Then, we determine a semantic distance between the categories of these neighbors and use it to answer the question if I_i, I_j show objects of the same class (fig. 6).

The motivation behind our distance is that we expect ImageNet to make connections between different instances that are not visually apparent (and thereby improve classification results). Consider three cases: (i) If I_i and I_j are visually very similar, they have very similar neighbors with a small semantic distance. In this case, both a visual distance and our semantic distance correctly classify the image pair. (ii) If I_i and I_j are visually quite dissimilar, but from the same class (e.g. two images of the basic-level category “car”, but one is a station wagon and the other a racing car), their neighbors are in different, but semantically related categories with low semantic distance. In this case, the semantic distance classifies the pair correctly while the visual distance does not. (iii) If I_i and I_j are not in the same class (e.g. a horse and a car), their neighbors will be in unrelated categories and thus the semantic distance will be high.

We name our method the *ImageNet Distance* and describe below how it uses visual nearest neighbors in ImageNet to infer a semantic distance between two images.

Below we detail two variants of the ImageNet Distance, which differ in how they compute the semantic distance between the neighbors. In sec. 5.3 we experimentally compare the ImageNet Distance to three purely visual distances (one simple direct distance and two trained for this task, sec. 5.2).

ImageNet Distance based on Jiang-Conrath ($D_{\text{IN}}^{\text{JC}}$). Given two input images I_i, I_j , we determine their k nearest neighbors $\mathcal{N}_i, \mathcal{N}_j$ in ImageNet. The ImageNet Distance based on JC measures the sum of the JC semantic distance between the categories of all pairs of neighbors $(n_i, n_j) \in \mathcal{N}_i \times \mathcal{N}_j$

$$D_{\text{IN}}^{\text{JC}}(I_i, I_j) = \sum_{n_i \in \mathcal{N}_i} \sum_{n_j \in \mathcal{N}_j} D^{\text{C}}(\mathcal{S}(n_i), \mathcal{S}(n_j)) \quad (5)$$

where $\mathcal{S}(n)$ denotes the category of neighbor n . Note how this approach allows to use any semantic distance between categories to derive a semantic distance between images.

ImageNet Distance based on Category Histograms ($D_{\text{IN}}^{\text{CH}}$). As above, we determine the k nearest neighbors $\mathcal{N}_i, \mathcal{N}_j$ of I_i, I_j in ImageNet. Then we compute the histogram h_i of the categories of the neighbors in \mathcal{N}_i . We repeat the operation for \mathcal{N}_j , giving h_j . These Category Histograms (CH) capture the category distribution of the neighbors of I_i and I_j . If I_i, I_j are from the same category, then we expect their neighbors to be distributed over the same categories. Therefore, we define our ImageNet Distance based on CHs to be the χ^2 distance between h_i, h_j

$$D_{\text{IN}}^{\text{CH}}(I_i, I_j) = \sum_{\mathcal{S}} \frac{(h_i(\mathcal{S}) - h_j(\mathcal{S}))^2}{h_i(\mathcal{S}) + h_j(\mathcal{S})} \quad (6)$$

where $h_i(\mathcal{S})$ is the number of neighbors of I_i in category \mathcal{S} .

5.2. Max-Margin Visual Distance Learning

As an alternative against which to compare our ImageNet Distance, we learn a purely visual distance using a max-margin formulation similar to [14, 41]. This distance $D_w(x_i, x_j)$ compares the visual descriptors x_i, x_j of the input images I_i, I_j . We focus on distances D_w defined as a weighted sum over component-wise differences

$$D_w(x_i, x_j) = \sum_d w_d \Delta(x_i^d, x_j^d) \quad \text{with} \quad (7)$$

$$\Delta(x_i^d, x_j^d) = \begin{cases} |x_i^d - x_j^d| & \Rightarrow \text{weighted } L_1 \\ (x_i^d - x_j^d)^2 & \Rightarrow \text{weighted } L_2 \\ \frac{(x_i^d - x_j^d)^2}{x_i^d + x_j^d} & \Rightarrow \text{weighted } \chi^2 \end{cases}$$

If $w_d = 1$ for all d , then $D_w(x_i, x_j)$ is simply the L_1, L_2 , or χ^2 distance. We aim at training w such that the distances for training pairs (x_i, x_j) of objects of the same class are smaller by a margin than the distances for pairs (x_k, x_l) of objects of different classes

$$D_w(x_i, x_j) \leq b - 1 < b + 1 \leq D_w(x_l, x_k) \quad (8)$$

$$\forall i, j, k, l \text{ with } c_i = c_j \text{ and } c_l \neq c_k$$

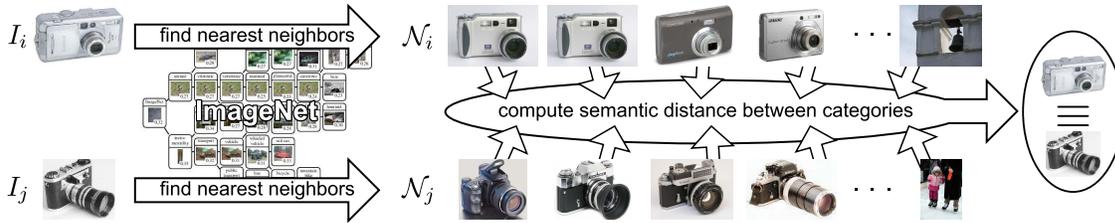


Figure 6: **Scheme of the ImageNet Distance.** For the input images I_i and I_j we search the nearest neighbors \mathcal{N}_i and \mathcal{N}_j in ImageNet using visual distances. Then, we determine a semantic distance between the categories of the neighbors and use it to answer the question if I_i and I_j show objects of the same basic-level category.

Let $X_n = \Delta(x_i, x_j)$ be the difference vector for pair $n = (i, j)$ (this plays the role of a feature vector describing the pair). Let $y_n = 1$ if $c_i \neq c_j$ and $y_n = -1$ if $c_i = c_j$ be the corresponding class label for the pair. We can rewrite inequality (8) as

$$y_n(w^T X_n - b) \geq 1 \quad \forall n = (i, j) \quad (9)$$

which is the constraint for a typical two-class support vector machine. Therefore, we minimize

$$\min_{w, \xi} \frac{1}{2} \|w\|^2 + C \sum_n \xi_n \quad (10)$$

$$\text{s.t. } y_n(w^T X_n - b) \geq 1 - \xi_n \quad \forall n = (i, j)$$

where ξ_n are slack variables and C is a regularization parameter balancing between margin and slack terms. We use liblinear [9] to minimize eq. (10). Note how with L_1 as Δ and a polynomial kernel of degree 2, our formulation learns a Mahalanobis distance (not necessarily positive definite).

We also compare to the large-margin nearest neighbor (LMNN) distance of [41] using the authors' software⁸ with default parameters and feature vectors reduced to 100 dimensions using PCA as recommended in the software manual.

5.3. Experiments

We evaluate the proposed ImageNet Distances on a subset of the Caltech101 dataset [10] containing 10 random images from each of the 102 classes (which are basic-level categories, such as “dog” and “car”). We randomly split the classes into two sets of 51 and use them as training and test sets (2-fold cross-validation). We learn the parameters of all distance functions on the training set (using the class labels) and evaluate them on the test set (with unknown class labels). In each set there are 129'795 pairs of images, 1.8% (2295) of them showing objects of the same class (positive pairs) and the others objects of different classes (negative pairs). Performance of classification into positive/negative pairs is measured by the area under the ROC curve (AUC).

For 54 of the 1020 images we use from Caltech101, we found a total of 125 duplicates or near-duplicates in ImageNet (the latter are images derived from Caltech101 images, e.g. by rescaling). We removed these duplicates from

ImageNet for this experiment, as they would artificially facilitate the task.

Visual Distance. We start by presenting the performance of purely visual distances. We first evaluate a direct χ^2 distance (not learned) between images, for 3 different descriptors (fig. 7a/table): GIST, LAB color histograms, and bag-of-visual-words histograms (sec. 4). We compute the χ^2 distance between all pairs of test images and then measure classification performance as the area under the ROC (AUC) curve. GIST outperforms the other descriptors, so we build the following experiments on it. We now evaluate learned distance functions between GIST descriptors. The LMNN distance [41] and the weighted L_1 and χ^2 distances all perform about equally well (fig. 7a). However, they do not improve significantly over the simpler direct χ^2 distance. This confirms that learning distance functions truly generic over classes is very difficult. As a side note, the L_2 distance performs significantly worse, confirming earlier findings that χ^2 is better suited for comparing image descriptors [21, 42].

It is important to notice that in our task the training and testing classes are completely *disjoint*. This makes the task significantly harder than having 5 images of each class for training and another 5 for testing (although it would involve the same set of images overall). In such an experiment the EER for GIST is 67.2% compared to 62.8% in our harder task.

ImageNet Distance. To find the nearest neighbors within the ImageNet Distances, we use χ^2 on GIST. We evaluate $D_{\text{IN}}^{\text{CH}}$ and $D_{\text{IN}}^{\text{JC}}$ using different numbers k of neighbors. For $D_{\text{IN}}^{\text{CH}}$, the AUC improves as k grows (fig. 7b) and converges at $k = 400$. A similar trend can be observed for $D_{\text{IN}}^{\text{JC}}$ (not in the figure), but its results are worse than $D_{\text{IN}}^{\text{CH}}$ (rows $D_{\text{IN}}^{\text{JC}}$ in the table). Importantly, $D_{\text{IN}}^{\text{CH}}$ with enough neighbors ($k > 100$) *outperforms all purely visual methods* (e.g. compare the dashed black line of $D_{\text{IN}}^{\text{CH}}$ to the red line of direct χ^2 on GIST).

Combining Visual and ImageNet Distance. We combine visual distances with ImageNet distances in a weighted sum trained by a linear SVM as in [16]. We combine all three direct visual distances (χ^2 on each of the three descriptors) and three $D_{\text{IN}}^{\text{CH}}$ distances obtained by changing the descrip-

⁸<http://www.cse.wustl.edu/~kilian/page3/page3.html>

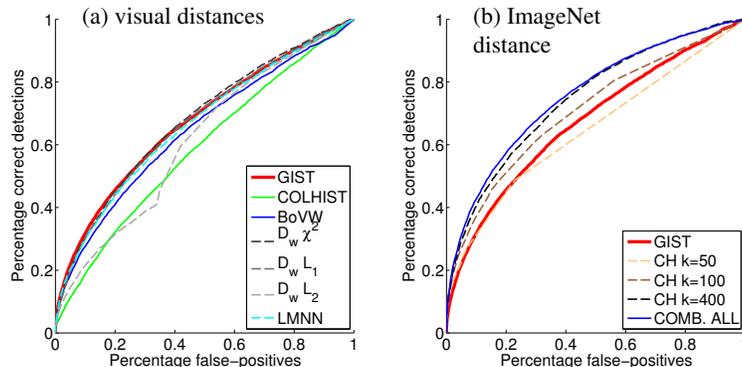


Figure 7: **Evaluation of the distance functions.** (a): performance of direct (GIST, COLHIST, BOW) and learned visual distances ($D_w\chi^2$, D_wL_1 , D_wL_2 , LMNN); (b): performance of ImageNet Distance D_{IN}^{CH} alone (CH) and combined visual and ImageNet distance (COMB.ALL). For reference, the performance of the best direct distance (GIST) is also given. The table shows the ROC AUCs and EERs of the various methods.

Method	AUC[%]	EER[%]
GIST direct	67.4	62.8
GIST learned χ^2	67.7	62.8
GIST LMNN [41]	66.4	61.7
D_{IN}^{JC} $k=400$	63.8	59.8
D_{IN}^{CH} $k=400$	74.7	67.2
COMB. ALL $k=400$	75.7	68.7

tor used to find the nearest neighbors. Interestingly, this combination brings a moderate improvement over one D_{IN}^{CH} and leads to our best result (blue line in fig. 7b; row COMB. ALL' in table).

6. Related Work

Large-scale image datasets [7, 27, 37] have been proposed for supporting the development of computer vision algorithms scalable to many images/classes. So far only a few applications have been investigated, such as retrieving specific objects [27, 38] and learning visual categories [6, 12, 29] and attributes [33]. In this paper instead, we employed ImageNet to study the relation between visual and semantic similarity (sec. 4). Most importantly, we investigated experimentally whether similarity measured through modern computer vision descriptors conveys semantic similarity. In cognitive science, analog questions were answered positively for the human visual system [15, 20, 30–32, 36]. The relationship between human and computer vision has been investigated for image retrieval, e.g. [34], and object recognition, e.g. [26]. Also the GIST descriptor has been designed to mimick human perception [25]. However, to the best of our knowledge, we presented the first large scale investigation of how well computer vision descriptors convey semantic similarity.

As a second main contribution, we presented the ImageNet distance to assess whether two images contain the same basic-level category (sec. 5). It is related to distance learning approaches in general [41] and for visual classification in particular [1, 13, 14, 22, 24, 28]. Typically a distance function is learned specific to one instance [22], or specific to one category [14], or more globally for a small set of 20-30 categories [1]. Nowak et al. [24] propose a distance to decide whether a pair of images contain the same object *instance* (not category). All these works tackle the problem based *purely on the two images*. Moreover, in these

works [1, 14, 22] the categories used for training the distance and for testing it are the same.

In contrast, our approach uses more than what available in the two images, as it exploits semantic similarity measured through ImageNet. Moreover, our goal is to build a *generic distance* to compare any two images, containing arbitrary unknown classes. In our experiments there is no overlap between training and testing classes. Such a general distance is useful for image retrieval [8, 13, 38], auto-annotation [17] and object recognition [2, 5, 10, 13, 23, 42].

In our ImageNet distance, we search for the nearest neighbors of the query images in ImageNet and compare their synset distributions to solve the original classification problem. This is a novel instance of the standard general strategy of using the output of a classifier as an intermediate representation for another classifier. It has been used, e.g. in e.g. *Classesmes* [39] for object recognition, to combine multiple kernels in [16], and in [35] where label histograms output by random forests are fed into SVMs. The closest work to our ImageNet distance is the image similarity measure of [40]. The output of 103 binary SVMs, each specific to one Flickr category, is used as a feature vector to compare two test images. In our work instead, the intermediate representation is formed by the *synset distribution* of the nearest neighbors of the test images. Moreover, we use a far larger dataset, more classes, and evaluate with disjoint training and test classes.

7. Conclusion

We experimentally investigated the relations between visual and semantic categories, and studied whether some assumptions taken in many computer vision approaches are valid. In particular, we have found that (i) the visual variability within a category grows with its semantic domain; (ii) visual similarity grows with semantic similarity; (iii) visual classes are separable across semantically defined bound-

aries.

As a second contribution, we presented a novel distance between pairs of images which assesses whether they show instances of the same basic-level category. It uses ImageNet as background data to make additional semantic connections beyond direct visual similarity. We showed experimentally that it outperforms purely visual distances.

References

- [1] B. Babenko, S. Branson, and S. Belongie. Similarity metrics for categorization: From monolithic to category specific. In *ICCV*, 2009.
- [2] H. Bay, A. Ess, T. Tuytelaars, and L. van Gool. SURF: Speeded up robust features. *CVIU*, 110(3):346–359, 2008.
- [3] I. Biederman. Recognition-by-components: A theory of human image understanding. *Psychological Review*, 94(2):115–147, 1987.
- [4] A. Budanitsky and G. Hirst. Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. In *NAACL*, 2001.
- [5] G. Csurka, C. Bray, C. Dance, and L. Fan. Visual categorization with bags of keypoints. In *ECCV Workshop on Stat. Learn. in Comp. Vis.*, 2004.
- [6] J. Deng, A. C. Berg, K. Li, and L. Fei-Fei. What does classifying more than 10,000 image categories tell us? In *ECCV*, 2010.
- [7] J. Deng, W. Dong, R. Socher, L.-j. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009.
- [8] M. Douze, H. Jegou, H. Sandhawalia, L. Amsaleg, and C. Schmid. Evaluation of GIST descriptors for web-scale image search. In *CIVR*, 2009.
- [9] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: a library for large linear classification. *JMLR*, 9:1871–1874, 2008.
- [10] L. Fei-Fei, R. Fergus, and P. Perona. One-shot learning of object categories. *PAMI*, 28(4):594–611, 2006.
- [11] C. Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.
- [12] R. Fergus, Y. Weiss, and A. Torralba. Semi-supervised learning in gigantic image collections. In *NIPS*, 2009.
- [13] A. Frome, Y. Singer, and J. Malik. Image retrieval and classification using local distance functions. In *NIPS*, 2006.
- [14] A. Frome, Y. Singer, F. Sha, and J. Malik. Learning globally-consistent local distance functions for shape-based image retrieval and classification. In *ICCV*, 2007.
- [15] P. Gärdenfors. *Conceptual spaces: The geometry of thought*. MIT Press, 2000.
- [16] P. V. Gehler and S. Nowozin. On feature combination for multiclass object classification. In *ICCV*, 2009.
- [17] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid. TagProp: discriminative metric learning in nearest neighbor models for image auto-annotation. In *ICCV*, 2009.
- [18] J. Hays and A. Efros. Scene completion using millions of photographs. In *SIGGRAPH*, 2007.
- [19] J. J. Jiang and D. W. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. In *International Conference Research on Computational Linguistics*, 1997.
- [20] G. Lakoff. *Women, Fire, and Dangerous Things: What Categories Reveal About the Mind*. U Chicago Press, 1987.
- [21] S. Maji, A. Berg, and J. Malik. Classification using intersection kernel support vector machines is efficient. In *CVPR*, 2008.
- [22] T. Malisiewicz and A. A. Efros. Recognition by association via learning per-exemplar distances. In *CVPR*, 2008.
- [23] T. Malisiewicz and A. A. Efros. Beyond categories: The visual memex model for reasoning about object relationships. In *NIPS*, 2009.
- [24] E. Nowak and F. Jurie. Learning visual similarity measures for comparing never seen objects. In *CVPR*, 2007.
- [25] A. Oliva and A. Torralba. Modeling the shape of the scene: a holistic representation of the spatial envelope. *IJCV*, 42(3):145–175, 2001.
- [26] D. Parikh and C. L. Zitnick. The role of features, algorithms and data in visual recognition. In *CVPR*, 2010.
- [27] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. In *CVPR*, 2008.
- [28] D. Ramanan and S. Baker. Local distance functions: A taxonomy, new algorithms, and an evaluation. In *ICCV*, 2009.
- [29] M. Rohrbach, M. Stark, G. Szarvas, I. Gurevych, and B. Schiele. What helps where and why? semantic relatedness for knowledge transfer. In *CVPR*, 2010.
- [30] E. Rosch. Cognitive representation of semantic categories. *Journal of Experimental Psychology*, 104(3):192–233, 1975.
- [31] E. Rosch. Principles of categorization. In E. Rosch and B. B. Lloyd, editors, *Cognition and Categorization*, pages 27–48. Lawrence Erlbaum Associates, 1978.
- [32] E. Rosch, C. Mervis, W. Gray, D. Johnson, and P. Boyes-Braem. Basic objects in natural categories. *Cognitive Psychology*, 8:382–439, 1976.
- [33] O. Russakovsky and L. Fei-Fei. Attribute learning in large-scale datasets. In *ECCV Workshop Parts & Attributes*, 2010.
- [34] A. Schwaninger, J. Vogel, F. Hofer, and B. Schiele. A psychophysically plausible model for typicality ranking of natural scenes. *ACM Trans. Appl. Perc.*, 3(4):333–353, 2006.
- [35] J. Shotton, M. Johnson, and R. Cipolla. Semantic Texton Forests for Image Categorization and Segmentation. In *CVPR*, 2008.
- [36] R. J. Sternberg. *Cognitive Psychology*. Wadsworth, 5th edition, 2008.
- [37] A. Torralba, R. Fergus, and W. T. Freeman. 80 million tiny images: A large dataset for non-parametric object and scene recognition. *PAMI*, 30(11):1958–1970, 2009.
- [38] A. Torralba, R. Fergus, and Y. Weiss. Small codes and large image databases for recognition. In *CVPR*, 2008.
- [39] L. Torresani, M. Szummer, and A. Fitzgibbon. Efficient object category recognition using classemes. In *ECCV*, 2010.
- [40] G. Wang, D. Hoiem, and D. Forsyth. Learning image similarity from flickr groups using stochastic intersection kernel machines. In *ICCV*, 2009.
- [41] K. Q. Weinberger, J. Blitzer, and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. In *NIPS*, 2005.
- [42] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: a comprehensive study. *IJCV*, 73(2):213–238, 2007.