Fine-Grained Object Recognition with Gnostic Fields

Christopher Kanan
Jet Propulsion Laboratory
California Institute of Technology
ckanan@caltech.edu

Abstract

Much object recognition research is concerned with basic-level classification, in which objects differ greatly in visual shape and appearance, e.g., desk vs duck. In contrast, fine-grained classification involves recognizing objects at a subordinate level, e.g., Wood duck vs Mallard duck. At the basic-level objects tend to differ greatly in shape and appearance, but these differences are usually much more subtle at the subordinate level, making fine-grained classification especially challenging. In this work, we show that Gnostic Fields, a brain-inspired model of object categorization, excel at fine-grained recognition. Gnostic Fields exceeded state-of-the-art methods on benchmark bird classification and dog breed recognition datasets, achieving a relative improvement on the Caltech-UCSD Bird-200 (CUB-200) dataset of 30.5% over the state-of-the-art and a 25.5% relative improvement on the Stanford Dogs dataset. We also demonstrate that Gnostic Fields can be sped up, enabling real-time classification in less than 70 ms per image.

1. Introduction

Fine-grained object classification refers to distinguishing among object categories at subordinate levels, e.g., bird species, domestic dog breeds, car models, and facial identity. Many real-world computer vision applications require fine-grained object categorization, e.g., automated surveillance systems that record the model of vehicles, and systems that classify fish species to measure the level of biodiversity in an ocean environment. With the exception of face identification, much computer vision research has focused on building systems for discriminating among basic-level categories, e.g., alligator vs automobile. Many of the best known benchmark datasets mostly contain basic-level objects, in which few visual features are shared among the majority of the categories, e.g., Caltech-101 [12], Caltech-256 [14], and PASCAL VOC [10]. Fine-grained classification is often harder than basic-level categorization because the differences among objects are more subtle, with fewer category-specific features. There has recently been a substantial amount of interest in non-face subordinate-level classification by the computer vision community (e.g., [3, 4, 5, 8, 23, 40, 41]).

In this work, we apply Gnostic Fields, a brain-inspired model of object classification, to the problem of fine-grained categorization. In 1967, Jerzy Konorski hypothesized that the brain contains regions existing near the top of the visual processing hierarchy that engage in the classification of mutually-exclusive categories [25], and he called these regions Gnostic Fields. In his theory, Gnostic Fields are comprised of competing gnostic sets, with one set per category. Each set contains a potentially redundant population of category specific gnostic neurons (units). Gnostic neurons coarsely encode particular views or properties of an object, while retaining a degree of tolerance to non-relevant changes in object appearance, scale, and location.

In the past decade, functional neuroimaging has yielded evidence for the existence of brain regions devoted to visual categorization. The fusiform gyrus has been implicated in subordinate classification of faces [22], and it exhibits selective activity when radiologists view scans [16] and when birders and car experts perceive birds and cars [13]. Neurons exhibiting characteristics similar to gnostic units have been found in many brain areas (see [15, 35] for reviews).

In [19], the first implementation of Konorski's Gnostic Field model was proposed, and it achieved state-of-the-art accuracy on image, sound, and electronic odor classification tasks. Unlike deep neural networks that learn features from pixel-patches (e.g., [27, 29]), Gnostic Fields operate on intermediate-level features. In [19] these intermediate-level features were dense SIFT descriptors, and in later work space-variant filters learned using independent component analysis were used [18].

In this paper, we first improve the Gnostic Field model in several ways, allowing it to scale to larger datasets. An overview of our model is given in Fig. 1. We then demonstrate that Gnostic Fields excel at two fine-grained recognition tasks: bird species categorization and dog breed categorization. Subsequently, we explore how the number of gnostic units and how chromatic/grayscale features influences both speed and accuracy, which was not explored in earlier work [19]. Finally, we use these results to show that Gnostic Fields can classify individual images in less than 70 ms without substantially impairing accuracy.

2. Related Work

Gnostic Fields can be interpreted as a kind of feedforward neural network, and they are related to both probabilistic neural networks [36] (PNNs) and radial-basis function networks (RBFNs) [30]. When a PNN or RBFN classifies an input, it compares it to a layer of pattern detection units to assess its similarity to the training data. Typically there is one pattern detection unit per training instance, but occasionally clustering has been used to acquire the pattern detection units (e.g., [31]). To combine information across the pattern detection units, PNNs sum the output of labeled pattern detection units and RBFNs apply linear regression to their output. Because images vary in size, algorithms for extracting dense descriptors generally produce a variable number of descriptors per image, and to use them with these models it would be necessary to somehow combine the descriptors into a single vector, e.g., by using spatial pyramid matching [28]. In contrast, Gnostic Fields innately expect a variable number of densely extracted descriptors as their input. Gnostic Fields assess each descriptor's similarity to coarsely encoded gnostic units (pattern detection units) for each category, with the outputs given by the most active units per category, analogous to the max-pooling operation in the Hierarchical Max (HMAX) model of object recognition [34]. This gives Gnostic Fields a degree of invariance to changes in object shape and appearance. Note that in the HMAX model, max-pooling is used to construct descriptors by pooling over similar features, e.g., features with the same orientation, whereas here it is used to measure the similarity to each category. After max-pooling, Gnostic Fields use a form of divisive normalization to modulate the network's activity, before finally combining information across all of the descriptors extracted from an image.

Beyond being feed-forward neural networks, Gnostic Fields share few similarities with recent deep neural network approaches to object recognition, e.g., [27, 29]. These algorithms are extremely powerful and instead of using hand-engineered descriptors they learn image features from natural images. Deep neural networks can even exhibit units with properties analogous to gnostic units as an emergent phenomenon [29]. However, to achieve good performance these algorithms require extremely large datasets, otherwise they may not generalize well. Gnostic Fields can be effective even with little training data [19], and they are also comparatively easy to implement as long as toolboxes for extracting dense descriptors, clustering, and learning linear classifiers are available.

In computer vision, the most similar model to Gnostic Fields is the Naive Bayes Nearest Neighbor (NBNN) model [2]. To classify an image, NBNN accumulates evidence from descriptors, with the evidence gathered per descriptor done using a nearest neighbor approach. Gnostic Fields use a nearest cluster center approach instead, allowing them to be much more computationally efficient so that they can be run on larger datasets. Both approaches abstain from the hard-binning done in the bag-of-words model.

With the exception of Gnostic Fields, neural networks have not been used on recent fine-grained recognition benchmarks, but there have been many other approaches used for this task, e.g., [3, 4, 5, 8, 23, 40, 41].

3. Model

3.1. Image Features

Gnostic Fields sit atop a sensory processing hierarchy. Implementation-wise, this means that distinctive high-dimensional features are used as their input. While self-taught learning [33] using independent component analysis or sparse coding could be used to learn visual features from natural images, we chose to use an engineered approach: dense Color SIFT (CSIFT) descriptors [37]. CSIFT descriptors are SIFT features extracted from an achromatic (luminance) channel and two opponent-color channels [37], and they can be quickly extracted from images. Dense CSIFT descriptors are extracted from densely sampled image locations, and they have been widely used in object recognition research (e.g., [19, 28]).

Prior to extracting CSIFT features, we resized each image to make its smallest dimension 128 pixels, with the other dimension chosen to preserve the image's aspect ratio. Because SIFT descriptors are sensitive to an image's gamma encoding [20], we applied a retina-like nonlinear brightness normalization procedure to the image [21]. This is given by

$$I'_c(z) = \frac{\log(\epsilon) - \log(I_c(z) + \epsilon)}{\log(\epsilon) - \log(1 + \epsilon)} \ , \tag{1}$$

where $\epsilon > 0$ controls the strength of the normalization and $I_c(z)$ is the image for RGB channel c at a particular location z. For each image, the ϵ that made the mean output value across all channels closest to 0.65 was chosen, although the value of ϵ was constrained to be between 10^{-6} and 0.5.

Subsequently, we converted the normalized image to CSIFT's Gaussian color space, and then extracted SIFT descriptors from each of the three color space channels. We used the dense SIFT implementation from the VLFeat toolbox [38], which was configured to use 11×11 spatial bins with a stride (step size) of 5 pixels. For each image, this configuration produced about 500–1500 128-dimensional

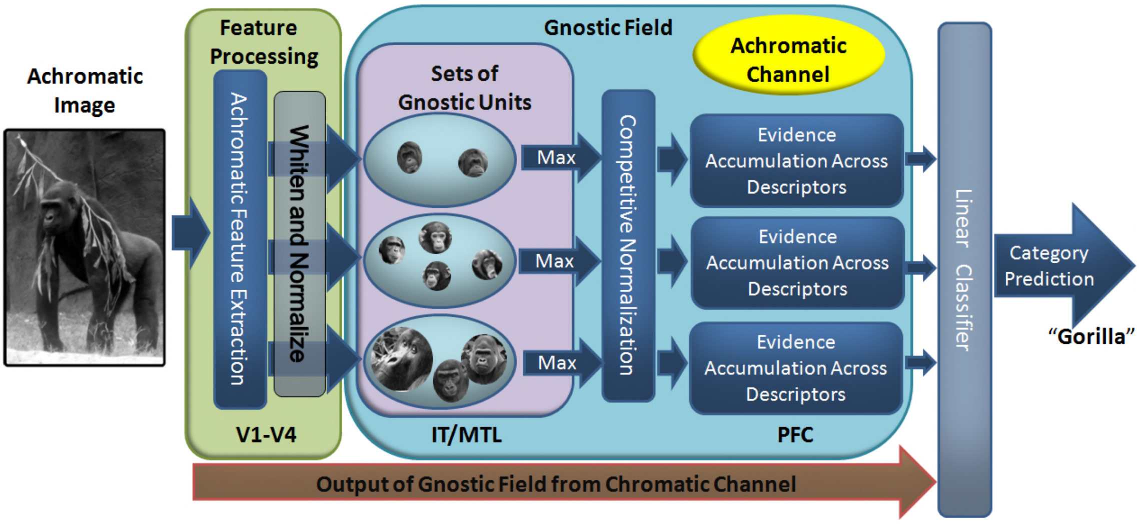
Achromatic Image
Feature Processing
Achromatic Feature Extraction
Whiten and Normalize
V1-V4
Gnostic Field
Sets of Gnostic Units
Max
Max
Max
IT/MTL
Competitive Normalization
Achromatic Channel
Evidence Accumulation Across Descriptors
Evidence Accumulation Across Descriptors
Evidence Accumulation Across Descriptors
PFC
Linear Classifier
Category Prediction
"Gorilla"
Output of Gnostic Field from Chromatic Channel

Figure 1. An example fine-grained Gnostic Field for categorizing among gorillas, chimpanzees, and orangutans. Achromatic image features are densely extracted from the image, and they are subsequently whitened and normalized. This is analogous to processing in early visual cortex. This information is sent to each gnostic set, with the units in the gnostic set for gorillas responding strongest. The output of each gnostic set is given by the most active gnostic unit. This activity is competitively normalized, which suppresses the output of the chimpanzee and orangutan sets. Evidence from all of the achromatic image descriptors is then accumulated. A linear classifier combines information from all of the achromatic gnostic sets as well as the chromatic gnostic sets (not shown) to predict the category.

feature vectors at different image locations for each of the three image channels. Only a single scale was used in our experiments. We then segregated the CSIFT descriptors into two channels: (1) the 128-dimensional achromatic channel alone and (2) a 384-dimensional channel formed by concatenating all three CSIFT channels together.

We augmented each descriptor $\mathbf{g}_{c,t}$ from channel c with topological information by appending a vector $\hat{\ell}_{c,t} = \frac{\ell_{c,t}}{||\ell_{c,t}||}$ to the descriptor, where $\ell_{c,t} = \left[x_t, y_t, x_t^2, y_t^2, 1\right]^T$ and (x_t, y_t) is the spatial location of $\mathbf{g}_{c,t}$ normalized by the image's dimensions (size) to be between -1 and 1. This yields $\hat{\mathbf{g}}_{c,t}$. We then learned whitening transformations with whitened PCA (WPCA) [1] for each of the two channels using the location augmented descriptors. WPCA learns a decorrelating transformation that normalizes the variance and can also be used for dimensionality reduction. The transformation is given by

$$\mathbf{U}_c = (\mathbf{D}_c + \xi\mathbf{I})^{-\frac{1}{2}} \mathbf{E}_c^T, \tag{2}$$

where $\mathbf{I}$ is the identity matrix, the columns of the matrix $\mathbf{E}_c$ contain the eigenvectors of the channel's covariance matrix, $\mathbf{D}_c$ is the diagonal matrix of eigenvalues, and ξ is a regularization parameter, with $\xi = 0.01$ in experiments. In [19] WPCA was applied directly to the training data, but this is infeasible with a large dataset. Instead, we applied WPCA to descriptors extracted from 584 images from the McGill color image dataset [32], which contains images of scenes. We only used the first 120 rows of $\mathbf{U}_c$, which yielded $\mathbf{W}_c$. Subsequently, the whitened descriptors are made unit length, allowing measurements of similarity using dot products [26]. The final 120-dimensional whitened and normalized descriptors $\mathbf{f}_{c,t}$ are given by

$$\mathbf{f}_{c,t} = \frac{\mathbf{W}_c \hat{\mathbf{g}}_{c,t}}{\|\mathbf{W}_c \hat{\mathbf{g}}_{c,t}\|}. \tag{3}$$

3.2. Gnostic Fields

We briefly provide the details necessary to implement Gnostic Fields here, but see [19] for additional information. A Gnostic Field for channel c is made up of K gnostic sets, with one set per category. Each gnostic set contains gnostic units that assess how similar the $\mathbf{f}_{c,1}, \ldots, \mathbf{f}_{c,T}$ descriptors from an image are to previous observations from that category. The output of a gnostic set for category k and from channel c is given by the unit in the set that is most similar to the descriptor (greatest dot product similarity), i.e.,

$$a_{c,k,t} = \max_j \left(\mathbf{v}_{c,k,j} \cdot \mathbf{f}_{c,t}\right), \tag{4}$$

where the max is taken over all of the $\mathbf{v}_{c,k,j}$ units (weight vectors) in the gnostic set. This max pooling step enables the gnostic set to vigorously respond to features matching the category's training data.

Spherical k-means [9] was used to learn the unit length $\mathbf{v}_{c,k,j}$ gnostic units for each of the $2K$ gnostic sets (K sets per channel) [19]. Spherical k-means is an unsupervised clustering algorithm for unit length data that learns unit length clusters [9]. The number of $\mathbf{v}_{c,k,j}$ units learned for a

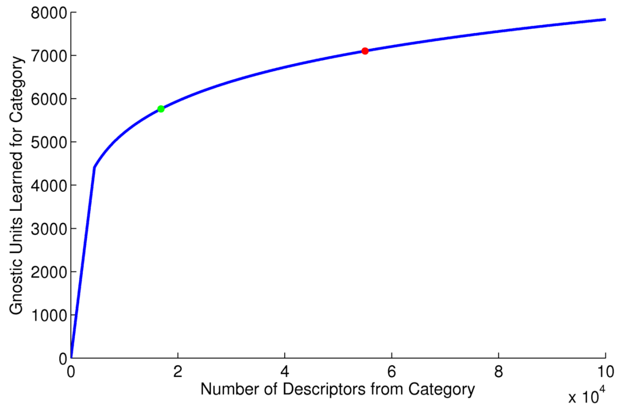

Figure 2. The total number of units learned for a gnostic set as a function of the number of descriptors extracted from images labeled with the gnostic set's category. In our main experiments, a median of 7100 units were allocated per gnostic set for the Stanford Dogs dataset (red dot) and 5760 units per gnostic set for the augmented CUB-200 dataset (green dot).

category k from channel c is given by

$$m\left(k,c\right)=\min\left(\left\lceil b\left(\log\left(n_{k,c}\right)+1\right)^{2}\right\rceil,n_{k,c}\right), \quad (5)$$

where $n_{k,c}$ is the total number of descriptors from category k and b regulates the number of units learned ($b=50$ in our main experiments, but see Section 4.5). This equation is plotted in Fig. 2, and it implements Konorski's idea that the number of gnostic units allocated to a gnostic set would increase with the amount of exposure to the set's object category, with fewer units being recruited as experience increases [25].

Gnostic Fields use inhibitive competition to suppress the output of the least active gnostic sets. For the K gnostic sets in channel c, this is implemented by attenuating their output using half-wave rectification [17], i.e.,

$$q_{c,k,t}=\max\left(a_{c,k,t}-\theta_{c,t},0\right), \quad (6)$$

with the threshold $\theta_{c,t}=\frac{1}{K}\sum_{k'}a_{c,k',t}$. The responses are then normalized using

$$\beta_{c,k,t}=\nu_{c,t}q_{c,k,t}, \quad (7)$$

with

$$\nu_{c,t}=\frac{\sum_{k'}q_{c,k',t}}{\left(K^{-1}+\sum_{k'}q_{c,k',t}^{2}\right)^{3/2}}, \quad (8)$$

acting as a form of variance-modulated divisive normalization (see [19]). This step has been previously reported to be crucial to achieving good image recognition accuracy using Gnostic Fields [19].

To accumulate categorical evidence across the entire image from each channel, Gnostic Fields simply sum the activity of the $\beta_{c,k,t}$ units across descriptors, i.e.,

$$\psi_{c,k}=\sum_{t=1}^{T}\beta_{c,k,t}. \quad (9)$$

Subsequently, the responses from all of these evidence accumulation units are combined across all categories and channels into a single vector $\mathbf{\Psi}$. This vector is then made mean zero and normalized to unit length.

A linear multi-category classifier is used to make the final categorical prediction, which was shown to improve performance by several percent in [19]. This allows less discriminative channels to be down weighted and it helps the model cope with confused categories. The model's predicted category is given by $\tilde{k}=\mathrm{argmax}_{k}\,\mathbf{w}_{k}\cdot\mathbf{\Psi}$, where $\mathbf{w}_{k}$ is the weight vector for category k. The $\mathbf{w}_{k}$ weights were learned with the LIBLINEAR toolbox [11] using Crammer and Singer's multi-class linear support vector machine formulation [6], with a low cost parameter (0.0001).

3.3. Implementation differences

There are several notable differences between our Gnostic Field implementation and that of [19]. In [19], WPCA was directly applied to the training features, which demands a very large amount of memory to handle big datasets. Here, we adopted a simple self-taught learning approach and applied WPCA to a separate dataset. No dimensionality reduction was done in [19]. Here, we used only a chromatic and an achromatic channel, whereas in [19] each CSIFT channel was used individually. We also changed how the number of gnostic units are allocated to each gnostic set, which will allow us to better control the number of units allocated (see Section 4.5).

4. Experiments

4.1. Bird species classification

The Caltech-UCSD Bird-200 (CUB-200) [39] is used for assessing methods for fine-grained object recognition. We used the 2010 version of the dataset, which is most widely used. It contains 6,033 images of 200 bird species, mostly from North America. There are 20–39 images per bird species. Example images are shown in Fig. 3A. Following the standard setting (e.g., [3, 8, 19, 23, 40]), the images were cropped to their bounding boxes. We assessed performance using the official train/test partitions, in which 15 images per category are used for training and the rest are used for testing. We also trained the model using a version of the training set augmented using horizontal reflections. Our results are given in Table 1. Our Gnostic Field implementation performs slightly better than the implementation of [19]. This indicates that our changes did not impair performance, even though we used two channels instead of three.

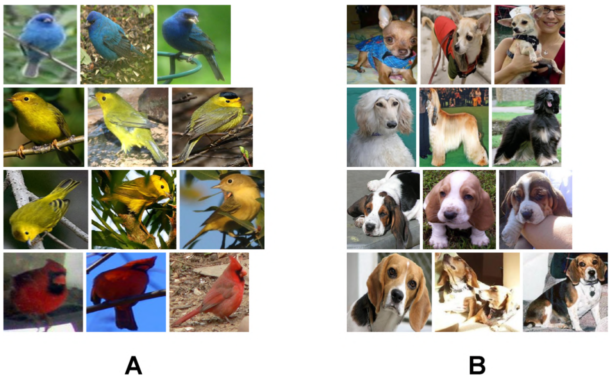

Figure 3. **A**. Example CUB-200 images from the Indigo Bunting, Wilson Warbler, Yellow Warbler, and Cardinal categories. The Wilson Warbler and Yellow Warbler share many traits. **B**. Example dog images from the Chihuahua, Afghan Hound, Basset Hound, and Beagle categories of Stanford Dogs. Note that some breeds do not have consistent coloration (e.g., Afghan Hounds), that the dogs are of various sizes and ages (i.e., full grown vs. puppy), and that multiple dogs can appear in a single image.

Using the augmented version of the training dataset, Gnostic Fields exceeded the best previously reported result [8], which also used an augmented version of the dataset, by 30.5% (in relative terms).

4.2. Dog breed classification

ImageNet [7] contains over 14 million high-resolution images belonging to almost 22,000 categories. The images were gathered from the Internet, and they were labeled by humans using Amazon's Mechanical Turk crowd-sourcing tool. A subset of ImageNet containing 120 different dog breeds has been used to create two fine-grained classification datasets: the ImageNet-2012 fine-grained classification challenge and Stanford Dogs [24]. We did experiments using both of these datasets. Example images are shown in Fig. 3B.

4.2.1 Stanford Dogs

The Stanford Dogs dataset [24] contains 20,580 images, with 148–252 images per category. Following the approach used by others (e.g., [24, 40]), we cropped the images to their bounding boxes, trained on 100 images per category using the official training partition, and tested on the remaining images. We did not augment the training dataset. Our results are given in Table 2. Using our full model, we achieved 47.7% accuracy, outperforming the previous best result of 38.0% [40] (a relative improvement of 25.5%).

4.2.2 ImageNet 2012 Fine-Grained Challenge

For the ImageNet 2012 Fine-Grained Challenge, the test labels are not publicly available. The official competition set contains 148–252 training images per category and 100,000 unlabeled test images, although only the 12,000 test images containing dogs are used by the evaluation server to measure performance.

Table 2. Mean per-class accuracy on the Stanford Dogs dataset using the official train/test partition. We assessed both the combined model (CSIFT) and a model using grayscale SIFT alone.

Model	**Accuracy (%)**
Gnostic Field (CSIFT) - Ours	47.7
Gnostic Field (SIFT) - Ours	40.3
Template Model (Edge Templates) [40]	38.0
Spatial Pyramid Matching (SIFT) [24]	22.0

We trained a Gnostic Field using the official competition training set, ran the model on the test images, and then uploaded our test image predictions to the competition server[1]. For this dataset, the quantitative measure of performance is the mean average precision (mAP), i.e., the mean of the average precision values calculated for each individual category. Our results are given in Table 3. Team ISI had the best model in the 2012 competition, and they used CSIFT, GIST, and RGB-SIFT features. Gnostic Fields exceeded this result using CSIFT features alone.

4.3. Running time of full model

Our experiments were done on a machine with an Intel Core i7-980X processor (introduced in 2010) and 24GB of RAM. Our machine also has a 6GB NVIDIA GeForce GTX TITAN graphics card, which was used to speed up dot products and matrix multiplications. All experiments were conducted using MATLAB R2013a.

For Stanford Dogs and CUB-200, classifying each image took 200–250 ms with the full model. This is sufficiently fast for Gnostic Fields to be used in many real-time or online classification tasks. Most of this time was dominated by the computations required by equations 3 – 9, with CSIFT feature extraction taking about 30 ms (10 ms per channel) per image. Learning the gnostic units required 31 s for the augmented CUB-200 dataset and 37 s per category for Stanford Dogs using our GPU-based implementation of spherical k-means. There are a number of ways the model could be further sped up, with the easiest being using fewer channels and/or fewer gnostic units per category. Both of these changes could potentially decrease accuracy, and we investigate this in the next two sections.

4.4. Individual channel performance

In our main results we used a linear classifier to fuse the output of achromatic and chromatic Gnostic Fields. We examined the benefit this provides by training linear classifiers

[1] The competition results are available at http://www.image-net.org/challenges/LSVRC/2012/results.html

Table 1. Comparison mean-per-class accuracy on CUB-200 [39]. Some reported results augment the size of the training set by horizontally flipping the training images. All reported results use color descriptors. Chance is 0.5%.

Model	Augmented	Accuracy (%)
Gnostic Field (CSIFT) - Ours	Yes	42.8
BubbleBank (SIFT + Color Histograms) [8]	Yes	32.8
Gnostic Field (CSIFT) - Ours	No	32.3
Gnostic Field (CSIFT) [19]	No	30.2
Template Model (Kernel Descriptors) [40]	No	28.2
TriCos (SIFT + Color Histograms) [4]	No	26.7
Multi-cue (CSIFT) [23]	No	22.4
Hierarchical Matching (SUN) [5]	No	19.2
Random Forest (CSIFT) [41]	No	19.2
MKL (HSV-SIFT + Geometric Blur) [3]	No	19.0

Table 3. Comparison results on ImageNet-2012 fine-grained classification challenge (task 3) with the top two teams. Note that we do not have a team name since our entry was uploaded in 2013.

Model	Team	mAP (%)
Gnostic Field (CSIFT) - Ours	N/A	36.461
Fisher Vectors (CSIFT + GIST + RGB-SIFT)	ISI	32.252
Fisher Vectors (SIFT + Color Features)	XRCE/INRIA	30.993

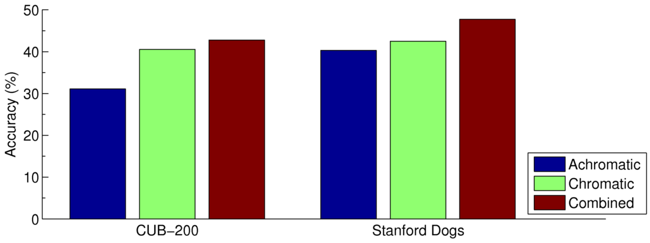

Figure 4. The results of the achromatic and chromatic Gnostic Fields alone, along with our main results using the combined approach, on CUB-200 and Stanford Dogs ($b = 50$ for these results). For both datasets chromatic features outperform achromatic features, although combining both achieves the highest accuracy.

on the output of each Gnostic Field individually for CUB-200 and Stanford Dogs. These results are given in Fig. 4. For CUB-200, our results use the augmented version of the dataset. For both datasets, a significant gain in performance is achieved by using the combined approach, although the improvement is relatively small for CUB-200 since color is very diagnostic for birds.

4.5. Gnostic neuron count, speed, and accuracy

Real-time object recognition is needed for many practical applications such as image search and robotic visual systems. To increase the speed of Gnostic Fields, the number of gnostic units learned should be minimized, since every descriptor is compared to every gnostic unit in each gnostic set. However, using too few units will degrade accuracy. To examine how the number of units allocated to a gnostic set influences accuracy and speed, we varied the value of the b parameter in equation 5, which linearly increases the number of units in each set. Because CUB-200 contains fewer images, and hence fewer descriptors, we focus on Stanford Dogs. We also limited this experiment to only the Chromatic channel because using both channels doubles the amount of time required while only giving a small, but significant, boost to performance (see Section 4.4).

Our results using only the Chromatic channel are shown in Fig. 5. As expected, accuracy was lowest when $b = 1$ and greatest when $b = 50$. However, using $b = 20$ images were classified in only 69.5 ms each (14 frames per second), while accuracy only decreased by 1.2% from when $b = 50$. This graceful scaling of accuracy and speed is a very desirable property for many real-world applications.

5. Discussion

In this paper we demonstrated that Gnostic Fields are effective at fine-grained object categorization by achieving state-of-the-art results on several subordinate-level classification benchmarks. We used only a single feature type at a single scale and performance would likely increase further if additional feature types were used. We also assessed the impact of the number of gnostic units in a gnostic set on the performance of Gnostic Fields, and we demonstrated that Gnostic Fields are suitable for real-time applications. While it is unlikely that our implementation's learning rules are implemented in the brain, the model uses neurally plausible operations during classification, e.g., dot products, max-pooling, and divisive normalization [19].

We are currently investigating how to make Gnostic Fields even faster during classification. In our implementation, each descriptor is compared to all of the units in every gnostic set. This means that without additional paral-

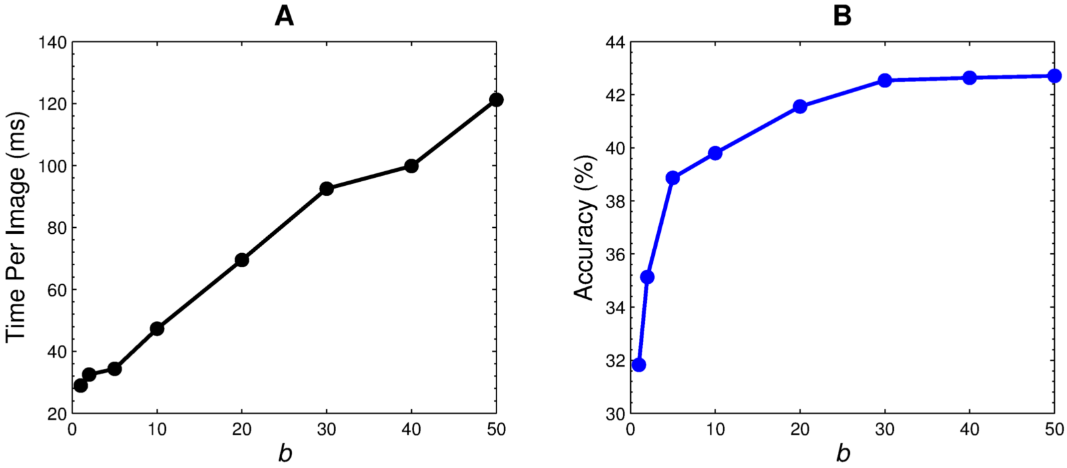

Figure 5. The influence of the the b parameter in equation 5 on speed and accuracy with the Chromatic channel alone. The number of units in each gnostic set increases linearly with b. **A**. As b increases the amount of time required per image increases approximately linearly. **B**. Mean per-class accuracy on the Stanford Dogs dataset as a function of b.

lelism, e.g., using a cluster with many machines or multiple GPUs, it is unlikely that Gnostic Fields could achieve real-time performance with very large datasets such as the ImageNet 22,000 object recognition challenge. One potential way to overcome this limitation would be to exploit basic-level subcategories, so that fine-grained gnostic sets are only evaluated if it is likely that the object belongs to its basic-level category.

Acknowledgements

The author would like to thank Gary Cottrell and Vicente Malave for feedback on earlier versions of this paper. This work was completed while the author was affiliated with the University of California San Diego. This work was supported in part by NSF Science of Learning Center grants SBE-0542013 and SMA-1041755 to the Temporal Dynamics of Learning Center.

References

[1] A. Bell and T. Sejnowski. The “independent components” of natural scenes are edge filters. *Vis Research*, 37:3327–3338, 1997.

[2] O. Boiman, E. Shechtman, and M. Irani. In defense of nearest-neighbor based image classification. In *CVPR*, 2008.

[3] S. Branson, C. Wah, B. Babenko, F. Schroff, P. Welinder, P. Perona, and S. Belongie. Visual recognition with humans in the loop. In *ECCV*, 2010.

[4] Y. Chai, E. Rahtu, V. Lempitsky, L. Van Gool, and A. Zisserman. TriCos: A tri-level class-discriminative co-segmentation method for image classification. In *ECCV*, 2012.

[5] Q. Chen, Z. Song, Y. Hua, Z. Huang, and S. Yan. Hierarchical matching with side information for image classification. In *CVPR*, 2012.

[6] K. Crammer and Y. Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *J Machine Learning Research*, 2:265–292, 2001.

[7] J. Deng, A. Berg, K. Li, and L. Fei-Fei. What does classifying more than 10,000 image categories tell us? In *ECCV*, 2010.

[8] J. Deng, J. Krause, and L. Fei-Fei. Fine-grained crowdsourcing for fine-grained recognition. In *CVPR*, 2013.

[9] I. Dhillon and D. Modha. Concept decompositions for large sparse text data using clustering. *Machine Learning*, 42:143–175, 2001.

[10] M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88:303–338, 2010.

[11] R. Fan, K. Chang, C. Hsieh, X. Wang, and C. Lin. LIBLINEAR: A library for large linear classification. *J Machine Learning Research*, 9:1871–1874, 2008.

[12] L. Fei-fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: an incremental bayesian approach tested on 101 object categories. In *CVPR*, 2004.

[13] I. Gauthier, P. Skudlarski, J. Gore, and A. Anderson. Expertise for cars and birds recruits brain areas involved in face recognition. *Nature Neuroscience*, 3:191–197, 2000.

[14] G. Griffin, A. Holub, and P. Perona. The Caltech-256. CNS-TR-2007-001, Caltech, Pasadena, 2007.

[15] C. Gross. Genealogy of the “grandmother cell”. *Neuroscientist*, 8:512–518, 2002.

[16] E. Harley, W. Pope, J. Villablanca, J. Mumford, R. Suh, J. Mazziotta, D. Enzmann, and S. Engel. Engagement of fusiform cortex and disengagement of lateral occipital cortex in the acquisition of radiological expertise. *Cerebral Cortex*, 19:2746–2754, 2009.

[17] D. Heeger. Half-squaring in responses of cat striate cells. *Visual Neuroscience*, 9:427–443, 1992.

[18] C. Kanan. Active object recognition with a space-variant retina. *ISRN Machine Vision*, 2013:138057, 2013.

[19] C. Kanan. Recognizing sights, smells, and sounds with gnostic fields. *PLoS ONE*, e54088, 2013.

[20] C. Kanan and G. Cottrell. Color-to-grayscale: Does the method matter in image recognition? *PLoS ONE*, 7:e29740, 2012.

[21] C. Kanan, A. Flores, and G. Cottrell. Color constancy algorithms for object and face recognition. *Lecture Notes in Computer Science (ISVC-2010)*, 6453:199–210, 2010.

[22] N. Kanwisher, J. McDermott, and J. Chun. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *J. Neuroscience*, 17:4302–4311, 1997.

[23] F. Khan, J. van de Weijer, A. Bagdanov, and M. Vanrell. Portmanteau vocabularies for multi-cue image representation. In *NIPS*, 2011.

[24] A. Khosla, N. Jayadevaprakash, B. Yao, and L. Fei-Fei. Novel dataset for ne-grained image categorization. In *First Workshop on Fine-Grained Visual Categorization, CVPR*, 2011.

[25] J. Konorski. *Integrative Activity of the Brain*. Univ. Chicago Press, Chicago, 1967.

[26] M. Kouh and T. Poggio. A canonical neural circuit for cortical nonlinear operations. *Neural Computation*, 20:1427–1451, 2008.

[27] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[28] S. Lazebnik, C. Schmid, and J. Ponce. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In *CVPR*, 2006.

[29] Q. Le, M. Ranzato, R. Monga, M. Devin, K. Chen, G. Corrado, J. Dean, and A. Ng. Building high-level features using large scale unsupervised learning. In *ICML*, 2012.

[30] J. Moody and C. Darken. Fast learning in networks of locally-tuned processing units. *Neural Computation*, 1:281–294, 1989.

[31] M. Musavi, W. Ahmed, K. Chan, K. Faris, and D. Hummels. On the training of radial basis function classifiers. *Neural Networks*, 5:595–603, 1992.

[32] A. Olmos and F. Kingdom. A biologically inspired algorithm for the recovery of shading and reflectance images. *Perception*, 33:1463–1473, 2004.

[33] R. Raina, A. Battle, H. Lee, B. Packer, and A. Ng. Self-taught learning: Transfer learning from unlabeled data. In *ICML*, 2007.

[34] M. Riesenhuber and T. Poggio. Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2(11):1019–1025, 1999.

[35] A. Roy. An extension of the localist representation theory: grandmother cells are also widely used in the brain. *Frontiers in Psychology*, 4(300), 2013.

[36] D. Specht. Probabilistic neural networks. *Neural Networks*, 3:109–118, 1990.

[37] K. van de Sande, T. Gevers, and C. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Trans Pattern Anal Machine Intell*, 32:1582–1596, 2010.

[38] A. Vedaldi and B. Fulkerson. VLFeat: An open and portable library of computer vision algorithms, 2008. Available: http://www.vlfeat.org. Accessed: 1 June 2012.

[39] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. CNS-TR-2010-001, Caltech, Pasadena, 2010.

[40] S. Yang, L. Bo, J. Wang, and L. Shapiro. Unsupervised template learning for fine-grained object recognition. In *NIPS*, 2012.

[41] B. Yao, A. Khosla, and L. Fei-Fei. Combining randomization and discrimination for fine-grained image categorization. In *CVPR*, 2011.