Hierarchical Classification of Images by Sparse Approximation

Byung-soo Kim[1]
bsookim@umich.edu

Jae Young Park[1]
jaeypark@umich.edu

Anna C. Gilbert[2]
annacg@umich.edu

Silvio Savarese[1]
silvio@eecs.umich.edu

[1] Department of EECS
University of Michigan
Ann Arbor, USA

[2] Department of Mathematics
University of Michigan
Ann Arbor, USA

Abstract

Using image hierarchies for visual categorization has been shown to have a number of important benefits including a significant gain in efficiency (e.g., logarithmic with the number of categories [11, 18]) or the construction of a more meaningful distance metric for image classification [19] (Fig. 1). However, a critical question still remains unanswered: would structuring data in a hierarchical sense also help classification accuracy? In this paper we address this question and show that the hierarchical structure of a database can indeed be used successfully to enhance classification accuracy using a sparse approximation framework. We propose a new formulation for sparse approximation where the goal is to discover the sparsest path within the hierarchical data structure that best represents the query object. Extensive quantitative and qualitative experimental evaluation on a number of branches of the Imagenet database [7] as well as on the Caltech-256 [11] demonstrate our theoretical claims and show that our approach produces better hierarchical categorization results than competing techniques.

1 Introduction

Recent advances in computer vision and machine learning permit the design of recognition methods that can classify images into large number of visual categories (typically, hundreds) [6, 8, 12, 15]. In a current paradigm for image categorization, image classes are organized in a flat structure and the problem is to discover the class (among all those in the flat structure) that best represents (in term of a distance function) the visual content of a given query image.

Recently, researchers have explored the idea of organizing visual data in a hierarchical structure rather than in a flat one. This paradigm addresses some of the limitations of flat representation: i) it allows to a significant gain in efficiency, typically logarithmic with the number of categories [11, 16, 18]; ii) it enables the construction of a more meaningful distance metric for image classification; iii) arranging visual data in a hierarchical structure echoes the way how humans organize data [16, 19]. A critical question, however, still remains unanswered: would structuring data in hierarchical sense also help classification accuracy? Currently there is no definite answer to that question. For instance, top-down classification schemes (applied on hierarchical structures) such as [11, 18] have produced

BMVC 2011 http://dx.doi.org/10.5244/C.25.106

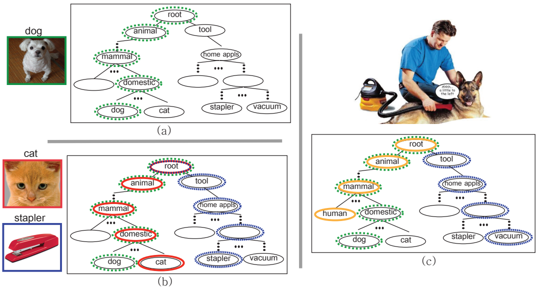

Figure 1: (a) Organizing images in a hierarchical structure (tree) enables more descriptive methods for characterizing images: the image of a dog can be described by class labels associated to each node of the (green) path in the tree. (b) Misclassifying a dog with a cat is not as bad as misclassifying a dog with a stapler. If data are organized in a tree, it is possible to relate object classification errors with objects with locations in the tree. For instance dog, cat and stapler categories are associated with the green, red and blue paths (respectively) in the tree. The error in misclassifying a dog with a cat can be measured as the Hamming Distance (HD) between the corresponding paths. HD captures the similarity between two paths in the tree (see Sec.3 for details). The HD is 1 in this case. Note that misclassifying a dog with a stapler leads to a larger HD (that is, 5). (c) It is desirable to classify multiple objects at the same time. If an image contains a dog, a human and a vacuum, our algorithm can discover three paths (green, orange and blue respectively) in the tree.

inconclusive evidence as to whether hierarchy has a beneficial effect on classification accuracy. Classification methods based on Hierarchical Support Vector Machines can be used to trade off accuracy against speed [11] or employed to increase classification accuracy as originally proposed in [22] and utilized for image classification in [2]. Although methods proposed by [2] have shown promising results, they are computationally very demanding as the number of categories becomes larger than 30∼50. Finally, methods based on combining models from different levels of the hierarchy [24] have also shown positive signal but are yet to be validated on deeper and larger hierarchical structures.

In this paper we address the issues discussed above and show that the hierarchical structure of a database can be successfully used to enhance classification accuracy using a sparse approximation framework. The key idea is to introduce a distance function that takes into account the hierarchical structure of the visual categories (Fig. 1) and define two images to be similar if they share a similar path in the hierarchy. We show that this distance function (or similarity metric) is equivalent to the Hamming distance (HD) for vectors that encode the hierarchy. This allows to cast the categorization problem as the one of discovering the category in the tree structure that has the smallest HD from the query category label. We solve this problem via sparse approximation and introduce a new formulation of the sparse approximation problem which we call *hierarchical* sparse approximation. In the typical sparse approximation problems, [6, 21, 23], a query image can be identified as the sparsest representation over the set of training images [23] or basis functions [17] for all object classes; that is, the sparsest solution is one (or a combination of a few) image out of *all possible* images in the dataset. We call this the *flat* sparse approximation problem. The key novelty of our approach relies on the idea that the sparse representation is not constructed over a *flat* structure of object classes (as in the classic sparse sensing problem) but rather by enforcing that the solution must be one (or a combination of a few) path out of all possible paths on a given hierarchy of object classes (training set). Moreover, classification accuracy is measured in hierarchical sense (that is, by considering the HD between the query path and the ground truth one). Since our method relies on the sparsity of the representation, our approach is suitable for large scale classification problems; i.e., the conditions underlying the sparsity assumptions are best verified when the dataset is large and distribution of visual categories is diversified. In this work we present sufficient conditions under which our hierarchical sparse formulation can be used with success and small error bounds are guaranteed. Furthermore, a crucial property of our classification framework is that it is capable of classifying multiple object categories at the *same time* if more than one (dominant) object appears in the query image (Fig. 1 (c)).

We have carried out extensive quantitative and qualitative experimental evaluation on a number of branches of the Imagenet database [7] as well as Caltech-256 [12]. Each branch comprises hundreds of visual categories organized in the hierarchical structure. All the experiments demonstrate that our hierarchical approximation framework yields better hierarchical classification accuracy over *flat* sparse approximation. Evaluation was carried out by comparing average precision measured in terms of HD as well as by measuring the actual classification accuracy at each level of the hierarchy. Our method achieves a performance increase ranging from 5% to 10% for the most critical levels of the hierarchy. Additional experiments on multi-category classification also show very promising results.

The rest of this paper is organized as follows. In Section 2, we will briefly review how sparse approximation can be applied to image classification problem. The formal definition of hierarchical classification and our proposed embedding is provided in Section 3. A number of experiments are performed to validate our scheme in Section 4. Finally, we summarize our work in Section 5.

2 Image Classification using Sparse Approximation

In this section, we describe our image representation and introduce the basic formulation of the flat image classification problem based on sparse approximation. We assume a database of images is available. Furthermore we assume that such a database comprises a large number of categories and each category has a large number of image instances. We assume that each image has a dominant object instance with some level of background clutter as in Caltech-256 [12] or ImageNet [7]. In classification, we assume that the query image (with unknown category label) contains one (or multiple) dominant object(s) whose category label is represented by the dataset. Of course, the query object instance itself is not included in the dataset. The classification problem can be solved by seeking, among all the images (object instances) in the database, the one that is closest to the query object(s). The category such image belongs to is the classification result. If the query image contains multiple dominant objects, the classifier must return multiple category labels associated to all of the dominant objects in the query image.

Object representation and distance function. Assessing whether an image is "close" to another one relies on the construction of a distance function which depends on the way the visual content of an image is represented. Following a common representation used in computer vision, we describe an image using a normalized histogram of codewords (i.e., the bag of words representation, also named BOW) [6] or, equivalently, a histogram capturing a spatial pyramid of codewords [10, 15]. In either cases, we denote such histogram by a vector x. Codewords are drawn from a learnt dictionary of vector quantized features as described in [6, 10, 15]. The size of the dictionary is denoted by K. Thus x is a column vector of size K, if we use a simple histogram of codewords to represent the image. Notice that other types of representations are also possible. The similarity between two images represented by x_i and x_j can be measured by computing the l_n norm distance between x_i and x_j, where n can be 1[1], etc. Similar images will have small distances.

Model matrix. Let us stack all the images in the database in a matrix H. Columns of H will correspond to column vectors x. Thus, H will be $K \times N$, where N is the number of images in the dataset. We call this matrix H the *flat model matrix*. Under the assumption that the database is large, any query image can be represented as a superposition of one or more images in the training data with small error e such that $x = Hm + e$, where $N \times 1$ vector m is called the *mixing vector* and consists of a few non-zero entries associated to the images in the database that contribute to represent the query image by superposition. Note that the error e captures background clutter and the intra-class variability. A similar representation was introduced in [23] and was shown to be suitable for face recognition problems. We argue this is a reasonable model for the generic object classification too as long as the training set is large enough (so it is highly likely that the query image is exceedingly close to at least one instance in the training set and, thus, m is sparse) and the image representation (BOW in our case) yields satisfactorily discriminative features for classifying object classes as demonstrated in [12, 15]. In order to further justify the model, we show empirical evidence that mixing vectors m are both fairly sparse and concentrated using a number of datasets. See [13] for more details.

Classification. Clearly m contains the information that allows us to estimate the class label of the query image. Therefore, the classification problem (*what is the object class?*) is recast into the problem of estimating the vector m (*where is a nonzero entry?*). Furthermore, this formulation allows us to discover multiple dominant object categories in the image. Suppose the image contains three objects as in Fig. 1 (c), then the query image may be expressed as a superposition of $s = 3$ training histograms and the nonzero entries of m will return the 3 classes appearing in x (i.e, dog, human and vacuum). Solving m is challenging because the system is under-determined ($N \gg K$) and has an infinite number of solutions. Because we postulate or seek a s-sparse mixing vector m, we can formulate this problem as a sparse approximation problem and seek to find the sparsest solution that best approximates (in ℓ_0 error) the observed instance.[2]

Problem 0. $\min \|m\|_0 \quad$ subject to $\quad \|Hm - x\|_2 \leq \varepsilon$.

Unfortunately, the above problem is an NP-hard problem in general (given an arbitrary matrix H and an arbitrary vector x). We can, however, solve this problem in polynomial time with appropriate geometric assumptions on H; if the maximum entry of the matrix $|H^*H - I|$, or the coherence[3] $\mu(H)$, of the matrix is small, then there are several algorithmic solutions. Let us assume for now that the training set contains the query image x. As proposed by [4, 23], one method is to observe that Problem 0 is an optimization problem with a non-convex objective function and that a convex relaxation of this problem yields a problem which can be solved efficiently with standard optimization techniques [5],

Problem 1. $\min \|m\|_1 \quad$ subject to $\quad \|Hm - x\|_2 \leq \varepsilon$.

A second algorithmic approach is to use a greedy algorithm, one that identifies image instances iteratively, such as Orthogonal Matching Pursuit (OMP). See [21] and the references therein for details on this algorithm. In [13] we show that the coherence between individual images decreases as a function of their hierarchical distance; thus, while the overall coherence $\mu(H)$ is high, with high probability, the coherence between any two images is quite small and OMP can distinguish among these images and choose a representation close to

[1] or even the l_0 pseudo-norm that counts the number of places in which the vectors disagree

[2] The pseudo-norm $\|\cdot\|_0$ counts the number of non-zero entries in a vector.

[3] An equivalent definition of $\mu(H)$ is the maximum dot-product of different columns of H, $\mu(H) = \max_{i \neq j} |\langle H_i, H_j \rangle|$.

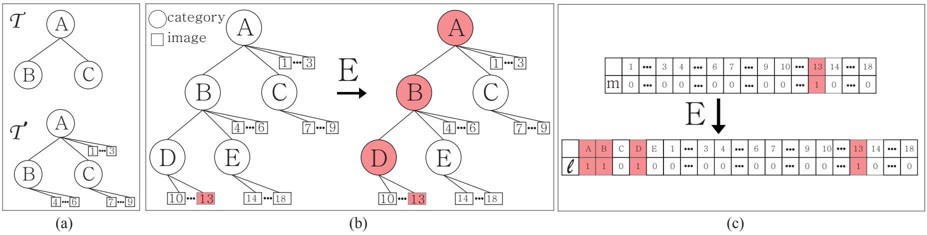

Figure 2: Visualization of the embedding. (a) Examples of $\mathcal{T}$ and $\mathcal{T}'$. (b,c) As a result of the embedding E, the flat mixing matrix m is mapped into l. In this example, when m shows a non-zero entry corresponding to image 13, the embedded l shows non-zero entries corresponding to image 13 as well as to its ancestors categories (nodes) A, B and D. These are on the path to the root from image 13.

the ground truth.

3 Hierarchical Classification with Sparse Approximation

While the model $x = Hm + e$ is reasonable and empirical evidence suggests that it is fairly accurate, it fails to take into account any hierarchical information amongst the classes. Furthermore, the error metrics for typical sparse approximation algorithms [5, 21] do not take into account structural relationships amongst the columns of H. Indeed, a small error in the mixing vector $\|\widehat{m} - m\|_2$ or in the reconstruction of the observation x does not necessarily guarantee hierarchical similarity between $\widehat{m}$ and m. For instance, suppose the ground truth label of a query image is “dog”. Assume two possible classification results are generated: “stapler” and “cat”. These two results would be associated to the same *flat* classification error $\|\widehat{m} - m\|_2$ if the model in $x = Hm + e$ were employed, whereas the classification error associated to “cat” would be smaller than that associated to “stapler” if the error function were defined in hierarchical sense (Fig. 1).

In this section, we assume that object categories are structured in a (rooted, labeled, recursive) tree $\mathcal{T}$ that reflects the semantic (parental) relationships among object categories. Note that each node of $\mathcal{T}$ contains all of the images representative of the visual category label associated to that node. A schematic illustration of such a data structure is given in Fig. 1 and 2. We define $\mathcal{T}'$, the data structure induced by the semantic tree, that contains two types of nodes, category labels and individual column vectors of H (images) (Fig. 2). It encodes the semantic relationship amongst the categories and the assignment of columns of H to those categories, but, unlike the tree $\mathcal{T}$, both categories and individual columns of H make up the nodes. A key contribution of our work is to introduce a suitable encoding matrix E that embeds the flat model matrix H into a hierarchical (tree) model matrix Φ and to show that the resulting hierarchical sparse approximation is solvable and appropriate for classification.

Hierarchical embedding. The encoding matrix E is constructed so as to map the mixing vector m into an embedded mixing vector $\ell = Em$, whose non-zero entries correspond to the paths in $\mathcal{T}'$ from the image to the root of the tree (Fig. 2). More concretely, we define E to be the $(N+C) \times N$ matrix that embeds a column of H and its path to the root in the tree $\mathcal{T}'$. Without loss of generality, we can permute the rows of E so that E has the following structure $E = [I \quad L]^T$ where I is the $N \times N$ identity matrix and the $C \times N$ matrix L consists of the hierarchical labels of each image. Each row of L corresponds to a category and each column to a training image; $L_{i,j} = 1$ if category i is on the path to the root from training image j. Each row encodes which training images are descendants of category j. Note that the length of ℓ is $N + C$. If we denote $E^\dagger$ the pseudo-inverse of E, then we define $\Phi = HE^\dagger$.

Hierarchical sparse approximation. The hierarchical embedding allows us to reformulate Problem 1 as a hierarchical sparse approximation problem and find a solution for ℓ given x:
Problem 2. $\min \|\ell\|_1 \quad$ subject to $\quad \|\Phi\ell - x\|_2 \leq \varepsilon$

Unlike the original sparse approximation problem, in this problem, the sparsity pattern of the vector ℓ is constrained to lie on a single path (or subtree) of the tree $\mathscr{T}'$. While the embedding $Em = \ell$ increases the number of non-zeros in ℓ (as compared to that of m), it also enforces a model that these non-zero entries must follow; they must lie on paths from individual columns of H to the root of the tree $\mathscr{T}'$. Because the sparsity of ℓ follows a model and Φ has more columns than rows, this problem has the structure of a model-based compressive sensing problem [1].

Problem 2 can be solved efficiently by a greedy algorithm called TREE-OMP [14], which is a special case of the more general algorithm MODEL-COSAMP [1], assuming that Φ satisfies a geometric condition, referred to as model-Restricted Isometry Property (model RIP). TREE-OMP is similar to the OMP algorithm with the additional step that for all non-zero components in the vector ℓ, the algorithm enforces that all the components that correspond to ancestors in the tree are non-zero. This constraint guarantees that the estimated solution $\widehat{\ell}$ corresponds to one (or more) physical path(s) in the tree.

Theoretical analysis. In this subsection, we show that the hierarchical embedding in Section 3 produces a matrix Φ that, on average, satisfies the model RIP. We also show that $\widehat{\ell}$, the output of TREE-OMP, is close to the ground truth embedded vector $\ell = Em$ not only in l_2 error, but, more importantly, in HD. This result enables the construction of a classification algorithm that we call SPARSE PATH SELECTION (SPS)(see below and [13]).

Theorem 1. *Given a normalized test image x ($\|x\|_2 = 1$) which is sd-sparse with background "noise" n, we can solve $\Phi\ell = x + n$ for the embedded mixing vector ℓ with* TREE-OMP. *After $T > \log(sd)$ iterations, the output vector $\widehat{\ell}$ has at most Td non-zero entries and satisfies*

$$\|\ell - \widehat{\ell}\|_2 \leq 2^{-T} + C\|n\|_2.$$

In addition, if the noise $\|n\|_2 \leq \sqrt{Td}(\eta - 2^{-T})$ is small enough compared to a learnt threshold η (See SPS algorithm), then $HD(\widehat{\ell}, \ell) = 0$; i.e., we correctly identify all the categories on the ground-truth hierarchical path.

Proof. First, we note that the embedded vector $\ell = Em$ follows a model-sparse pattern as defined in [1].

Lemma 1. *If m is a s-sparse vector, then $\ell = Em$ has a sparse tree structure; that is, it encodes a rooted tree with s leaves.*

Proof. The details of the proof are reported in [13]. □

Lemma 2. *The matrix Φ is well-approximated by an iid (sub-)Gaussian random matrix.*

Proof. The details of the proof are reported in [13]. □

From Lemma 1 and 2, we can conclude that Φ satisfies a model-RIP property [3]. Furthermore, we can use the result in [1] to conclude that after T iterations of TREE-OMP, the output $\widehat{\ell}$ contains at most Td non-zero entries and satisfies $\|\ell - \widehat{\ell}\|_2 \leq 2^{-T} + C\|n\|_2$. While the l_2 distance between two vectors is meaningful, it does not tell us how close the path(s) corresponding to the vector $\widehat{\ell}$ are compared to the ground-truth vector ℓ, it conflates the paths with the coefficients on those paths. The error bound tells us what the average error in $\widehat{\ell}$ is and, as long as it is below our learned threshold, $\frac{1}{\sqrt{Td}}(2^{-T} + \|n\|_2) < \eta$, we will not introduce spurious nodes in the path nor miss them and hence, $HD(\widehat{\ell}, \ell) = 0$. □

Sparse Path Selection Algorithm (SPS). After solving Problem 2, we obtain an estimate of the path ℓ in the hierarchical database associated to the query image. However, ℓ cannot be used as is for image classification. Ideally, the sparsest solution of Problem 2 should return a vector of "1" and "0" where the non-zero elements in ℓ allows to estimate the category labels of the query object as well as its parents. Unfortunately, this is not always the case and values between "0" and "1" can be also found because of the estimation noise. To solve this issue, we perform a post processing step. The idea is to introduce a threshold η and interpret as a positive response any value that is above such threshold (and as negative response, otherwise). Finding this threshold, however, is not trivial as it may be different if different datasets are used. Thus, in our experiments, we propose to automatically learn this threshold using a binary MAP estimator trained using a validation set. Such evaluation set is then removed from the dataset so as to avoid contamination during testing. Details of the SPS algorithm can be found in [13].

Classifying multiple categories. If the input vector x describes an image comprised of s categories, the mixing vector m is a s-sparse vector and the corresponding embedded mixing vector ℓ defines a subtree composed of s paths. Each of these paths is associated to one of the categories in x (Fig. 1). Thus, solving Problem 2 and obtaining an estimate $\widehat{m}$ of m allows us to *simultaneously* discover the presence of multiple categories in the image. However, one critical question must be addressed; how many categories s can we simultaneously handle until the conditions (i.e., sparsity, etc) underlying the solution of Problem 2 are violated? The bounds in [1] suggest that we need at least $O(sd)$ rows in the histograms, where d is depth of hierarchical tree. Section 4 gives empirical evidence that classification of multiple categories is possible using such procedure.

4 Experiments

In this section, we present quantitative and qualitative experimental results to validate our theoretical claims. We test our algorithm using different hierarchical databases. These are: i) two branches of the ImageNet [7] each comprising hundreds of categories; ii) The hierarchical Caltech-256 dataset [11]. We use different metrics to evaluate the performances of our algorithm: i) Overall average Hamming Distance (HD); ii) Average classification accuracy for each levels of the hierarchy. We benchmark our results using a competitive classification methods such as SRC, i.e., the sparse approximation technique introduced by [23]. Our experiments include classification of a single dominant object category as well as multiple categories. In each of the single category classification experiments we used 16 patches on a grid with step 8 pixels to generate SIFT descriptors. BOW histograms are constructed using 500 codewords generated from K-means clustering. Finally, we used SPH (Spatial Pyramid Histogram) up to the resolution level 4 to represent each image. In each experiment we sample (at most) 100 images for each node of the working database and use these for learning (i.e. to build the H matrix). For example, for the *domestic Animal* sub-tree of ImageNet, we collected about 21000 images for training. We sample an additional 10 images per node for testing.

ImageNet Subsets. ImageNet [7] is a hierarchical image database with $10{,}000{,}000$ images and over $10{,}000$ categories. It organizes different classes of images according to the WordNet [9] structure, and "IS-A" relationship exists between parents and children. In the experiments, we used two different branches from the ImageNet: *Domestic Animals*, and *Fruits*. These subsets are chosen so as to observe the effect of different numbers of categories (213, 320 respectively) and structures of the hierarchy (domestic animals has 7 levels whereas fruit has 6 levels) on the classification results.

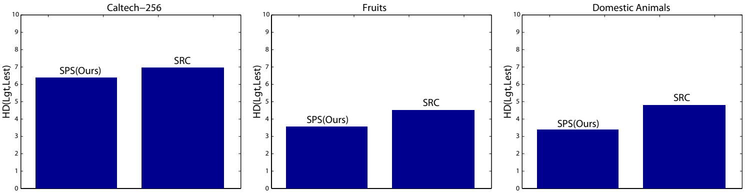

Figure 3: Average Hamming distance (HD) for different subcategories is drawn.

Hierarchical Caltech-256. Following [11, 12], the Caltech-256 is rearranged in a hierarchy according to best matches in the WordNet. In this Hierarchical Caltech-256, all images are associated to a leaf node, hence there are no images in the internal nodes. We study the properties of this dataset in [13]. The analysis suggests that solving Problem 2 gives us an answer that is close to the ground truth with high probability.

Benchmarks. The sparse approximation technique introduced by [23] (SRC) is used. We use problem 1 (Sec.2) to find the solution m via sparse approximation (similarly to [23]). We use the post-processing procedure in [23] to estimate the final class label. Notice that this method does not exploit the hierarchical structure of the database and "see" the database as flat. Notice that SRC returns a single class label (not a path in the tree) which can be used to form the mixing vector m_{SRC}. In order to compare SRC results with ours, we embed m into its corresponding path $\ell_{\mathrm{SRC}} = Em_{\mathrm{SRC}}$. Notice that classifying ℓ correctly is as challenging as classifying m correctly since we don't know in advance the depth of the ground truth path.

Hierarchical Similarity Verification. In this section, we show classification results in terms of HD (which is a natural distance function to compare the similarity of two paths in a tree). Thus, if the ground truth path and the estimated path are similar, the HD will be small. In Fig. 3 we show average HD between ground truth paths and estimated path for all our testing images using our approach (SPS). In the same figure we also report the average HD distance between ground truth path and path estimated by SRC (i.e., ℓ_{SRC}). Note that the HD associated to our approach is systematically smaller for all the datasets. This result supports our argument that the proposed framework yields smaller HD bounds. Also, notice that when the hierarchical structure is relatively flat, the effect of encoding (and thus the advantage of using our framework) becomes less significant.

Effect on Different Hierarchy Levels. HD returns a global measurement of path similarity regardless of the level and position in the tree. In this experiment we explore the performance of our framework at different levels of the tree. Figure 4 plots the accuracy versus the levels of the hierarchy for different datasets (see caption for details). Notice that the root node is always classified correctly. As we go down toward the bottom of the tree, the likelihood of classifying nodes correctly becomes smaller and smaller. Also, note that this graph is always monotonically decreasing because whenever the estimation of the child category is correct, parent category estimation is correct too. When the hierarchical level is low, the performance of our SPS is similar to SRC. Interestingly, the plot shows that two algorithms yield equivalent performances in classifying images belonging to the leaf nodes. However, when the hierarchical level increases, the gap between our SPS and SRC become much larger. This demonstrates the ability of our method to yield higher rates in classifying ancestors of the query object category, thus, yielding a smaller error in hierarchical sense. Anecdotal examples of paths returned by our SPS algorithm compared with those returned by SRC are shown in Fig. 5. Note that estimated parent nodes returned by SRC are much less accurate than those returned by SPS. Paths are reported in text format.

Multiple Category Classification. We report anecdotal examples demonstrating that our

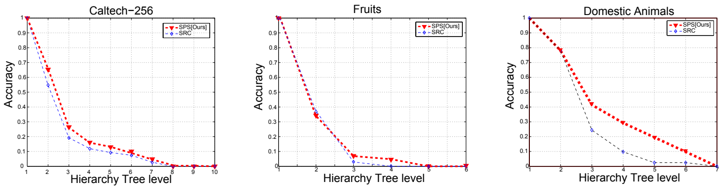

Figure 4: Average accuracy of classification for different hierarchical levels on three different categories, Caltech-256, Fruits, Domestic animals. Each plot captures the average number of correctly estimated nodes (categories) for each level (x-axis) for all testing images. A node j is estimated correctly if the ground truth path evaluated at j is equal to the estimated path at j for a given test image. Consider the example in Fig. 2 (b). In this case, the accuracy is be calculated over 3 levels. Suppose the ground truth query image is 13 and the ground truth m is associated to class labels A,B,D. If the estimated m returns class labels A,B,E, the accuracy for three levels is $1,1,0$. If the estimated m returns the class labels A,C, the accuracy for three levels is $1,0,0$. If the estimated m just returns the class labels A, the accuracy for three levels is still $1,0,0$.

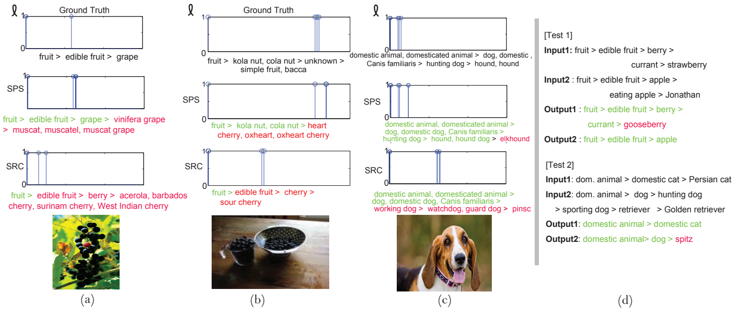

Figure 5: (a,b,c) Anecdotal examples of single object category recognition. The hierarchical path is estimated as nonzero entries in the encoded mixing vector ℓ. Note that the path estimated by SPS (ours) is closer to the ground truth path than that of SRC [23]. Green (red) indicates correct (incorrect) classification. (d) Anecdotal examples of multiple object category recognition obtained by our method.

framework is able to classify images containing multiple categories. In such examples the histogram representing the query image can be expressed as a superimposition of multiple object category histograms. So, as discussed in the technical section, our SPS method will return *multiple* paths – a path for each category in the query image. Examples in Fig. 5 show some successful cases. Paths are presented in text format. More examples along with numerical results showing how accurately the algorithm is capable to retrieve multiple images are reported in [13].

5 Conclusion

In this work, we introduced a novel framework for hierarchical classification using a new formulation of the sparse approximation problem. We demonstrated that the hierarchical structure of a large and complex database can indeed be used successfully to enhance classification accuracy. Experimental results on several large scale dataset were used to support our claims.

Acknowledgments

We acknowledge the support of a Google Research Award and the Gigascale Systems Research Center. The first author is partially supported by Korea Foundation for Advanced Studies. We wish to thank Anush Mohan for helpful feedback.

References

[1] R.G. Baraniuk, V. Cevher, M.F. Duarte, and C. Hegde. Model-based compressive sensing. *Information Theory*, 2010.

[2] A. Binder, M. Kawanabe, and U. Brefeld. Efficient Classification of Images with Taxonomies. In *ACCV*, 2009.

[3] T Blumensath and M Davies. Sampling theorems for signals from the union of linear subspaces. *Information Theory*, Jan 2007.

[4] E.J. Candès, J.K. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. In *Communications on Pure and Applied Mathematics*, 2006.

[5] S.S. Chen, D.L. Donoho, and M.A. Saunders. Atomic decomposition by basis pursuit. In *SIAM review*, 2001.

[6] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *Workshop in ECCV*, 2004.

[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. Everingham, A. Zisserman, C. Williams, and L. Van Gool. The PASCAL visual object classes challenge 2006 (VOC2006) results. In *Workshop in ECCV*, 2006.

[9] C. Fellbaum et al. *WordNet: An electronic lexical database*. MIT press Cambridge, MA, 1998.

[10] K. Grauman and T. Darrell. The pyramid match kernel: Discriminative classification with sets of image features. In *ICCV*, 2005.

[11] G. Griffin and P. Perona. Learning and using taxonomies for fast visual categorization. In *CVPR*, 2008.

[12] G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. 2007.

[13] B. Kim, J.Y. Park, A.C. Gilbert, and S. Savarese. Hierarchical classification of images by sparse approximation. In *Technical Report, CSPL, Dept. EECS, University of Michigan*, 2011.

[14] C. La and M.N. Do. Tree-based orthogonal matching pursuit algorithm for signal reconstruction. In *ICIP*, 2006.

[15] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.

[16] L.J. Li, C. Wang, Y. Lim, D.M. Blei, and L. Fei-Fei. Building and using a semantivisual image hierarchy. In *CVPR*, 2010.

[17] J. Mairal, F. Bach, J. Ponce, G. Sapiro, M. Wood, N. Hengartner, E. Matzner-Lober, L. Rouvière, T. Burr, N.V. Malyshkina, et al. Online learning for matrix factorization and sparse coding. 2009.

[18] M. Marszalek and C. Schmid. Semantic hierarchies for visual object recognition. 2007.

[19] S.E. Palmer. *Vision science: Photons to phenomenology*. MIT press Cambridge, MA., 1999.

[20] Robert Tibshirani. Regression shrinkage and selection via the lasso. In *Journal of the Royal Statistical Society, Series B*, 1994.

[21] J.A. Tropp. Greed is good: Algorithmic results for sparse approximation. In *Information Theory*, 2004.

[22] I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun. Support vector machine learning for interdependent and structured output spaces. In *ICML*, 2004.

[23] J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastry, and Y. Ma. Robust face recognition via sparse representation. 2009.

[24] A. Zweig and D. Weinshall. Exploiting object hierarchy: Combining models from different category levels. In *ICCV*, 2007.