Sparse Distance Learning for Object Recognition Combining RGB and Depth Information

Kevin Lai*, Liefeng Bo*, Xiaofeng Ren†, and Dieter Fox*†
*Department of Computer Science & Engineering
University of Washington, Seattle WA 98195, USA
†Intel Labs Seattle
Seattle WA 98105, USA

Abstract— **In this work we address joint object category and instance recognition in the context of rapid advances of RGB-D (depth) cameras [16, 3]. We study the object recognition problem by collecting a large RGB-D dataset which consists of 31 everyday object categories, 159 object instances and about $100,000$ views of objects with both RGB color and depth. Motivated by local distance learning where elementary distances (over features like SIFT and spin images) can be integrated at a per-view level, we define a view-to-object-instance distance where per-view distances are weighted and merged. We show that the per-instance distance, through jointly learning the per-view weights, leads to superior classification performance on object category recognition. More importantly, the per-instance distance allows us to find a sparse solution (through Group-Lasso), where a small subset of representative views of an object are identified and used, and the rest discarded. This not only reduces computational cost but also further increases recognition accuracy. We also empirically compare and validate the use of visual (i.e. RGB) cues and shape (i.e. depth) cues and their combinations.**

I. Introduction

Visual recognition of object categories and instances is a fundamental and challenging problem and a major focus of research for computer vision, machine learning, and robotics. In the past decade, a huge variety of features and algorithms have been proposed and applied to this problem, resulting in significant progress on object recognition capabilities, as can be seen on the steady improvements on standard benchmarks such as Caltech101 [6].

While most modern-day recognition benchmarks are constructed using Internet photos at the category level only, the goal of our work is to study the recognition problem at both the category and the instance level, on objects that we commonly use in everyday tasks. The ability to recognize objects at both levels is crucially important if we want to use such recognition systems in the context of specific tasks, such as human activity recognition or service robotics. For instance, identifying an object as a generic "coffee mug" or as "Amelia's coffee mug" can lead to substantially different implications depending on the context of a task.

In addition to category and instance level recognition, we want to enrich the recognition data by taking advantage of recent advances in sensing hardware. In particular, the rapidly maturing technologies of RGB-D depth cameras [16, 3] provide high quality synchronized videos of both color and depth, presenting a great opportunity of combining color- and depth-based recognition. Here, we use RGB-D cameras to collect a comprehensive object recognition dataset, which comprises 31 categories of everyday objects, 159 object instances, and about 100 views per instance, with both color and depth (Section IV). This dataset presents unique challenges for recognition, for instance: (1) how can we handle category and instance recognition together, (2) how can we integrate color and depth cues for recognition, and (3) how can we adequately handle a large number of views of the same object?

Given the extensive knowledge we have about individual cues for recognition,such as SIFT [13] for appearance and Spin-Image [11] for 3D shape, we seek to combine these cues and their metrics, and one successful line of work that does so is that of distance learning (e.g. [22, 21]), or in particular local distance learning [18]. Local distance learning has been extensively studied and demonstrated for object recognition, both for color images [8, 9, 14] and 3D shapes [12]. A key property of these approaches is that they can model complex decision boundaries by combining elementary distances.

Local distance learning, however, is not without issues. For our particular task, there are two main limitations to overcome: (1) existing formulations of local distance learning do not capture the relations between object categories and specific instances under them; (2) they provide no means for selecting representative views, or examples, of instances and thus become very inefficient if a large number of views are collected for each object.

In this paper we propose an approach of sparse instance-based distance learning: instead of learning per-view, or per-exemplar, distances, we combine local distances to define and optimize a *per-instance distance* I. By learning a distance function jointly for all views of an instance, our approach significantly outperforms view-based distance learning for RGB, Depth, and RGB+Depth recognition (this result can also be motivated as subclass classification [20, 5]). However, even more importantly, joint instance-based learning naturally leads to a sparse solution using Group-Lasso, where a sparse set of views of each instance is selected from a large pool of views. Thus, our approach is able to significantly sparsify the data set; discarding redundant views and speeding up classification. We show that the sparse solution does not decrease performance; on the contrary, higher accuracy is achieved by using a smaller

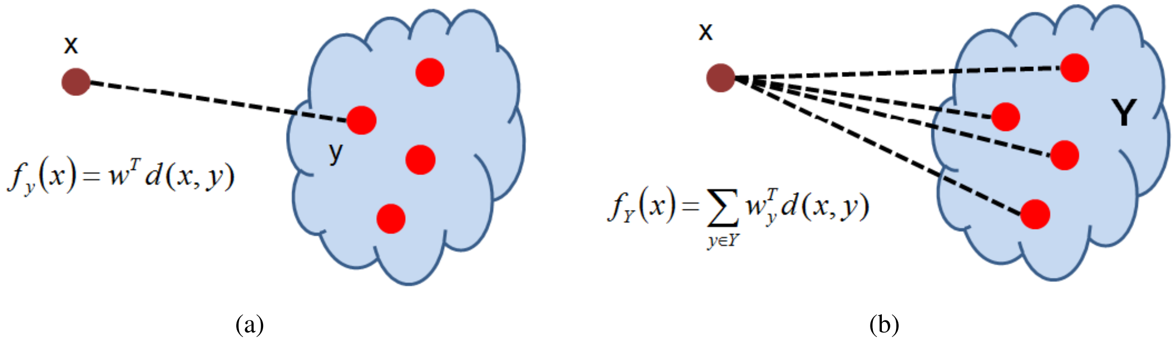

Fig. 1. In this work we study joint object category and instance recognition at a large scale through the use of per-instance distances. (a) Local distance learning uses a view-to-view distance, typically followed by a k-nearest neighbor rule. (b) We directly use the weighted average distance from an view x to an intance $\mathbf{Y}$ which consists of a set of views of the same object. This allows us to jointly learn the weights for each individual view y in the instance set, outperforming local distance learning. More importantly, this allows us to use Group-Lasso on $\mathbf{w}$ to select a sparse set of exemplars from a large pool of views.

but more representative set of exemplars for classification.

This paper is organized as follows. Our instance-based distance learning approach for object category and instance recognition is presented in Section 2. Then, in Section III, we introduce Group-Lasso sparsification, followed by experimental results and conclusions.

II. Learning Instance-Based Distances

In this section, we show how to learn instance-based distance functions for visual recognition tasks. In image classification, we are given a set of images Y and their corresponding label set t. The goal is to learn a classifier to recognize category and instance labels of images, or views, outside of the training set. One of the simplest methods to do this is to find nearest neighbors of the test view and make a prediction based on the labels of these nearest neighbors. In this section, we show how to improve this approach by learning an instance-based distance function. We start by considering a simple classification rule, *nearest instance classifier*, which labels incoming test images x using the label of the nearest instance (an extension to k-nearest instances is straightforward):

$$c_x = \operatorname*{argmin}_{i,j} \frac{1}{|Y_{ij}|} \sum_{y \in Y_{ij}} d(x, y) \tag{1}$$

Here, Y_{ij} denotes the set of views taken of the j-th instance of the i-th category. As can be seen, c_x is the object instance that appears, averaged over its views, most similar to the test image. $d(x, y)$ can be any distance function between views x and y. In this paper, we use the L_2 distance $d(x, y) = \|x - y\|^2$. The nearest instance classifier given in (1) can be used for both category and instance recognition: The index i provides the category and the index j gives the corresponding instance. Unfortunately, the nearest instance classifier can often perform poorly in practice due to the difficulties of finding a good distance measure. Instead, we now consider a significantly more powerful variant by learning an instance-based distance function for recognition.

In many problems there are multiple features available and the best performance is obtained by using all available information. To do so, we replace the scalar distance $d(x, y)$ between two views x and y by a vector $\mathbf{d}(x, y)$ of separate L_2 feature differences. The corresponding instance based distance function between example x and the j-th instance of i-th category Y_{ij} then can be written as

$$f(x, \mathbf{W}_{ij}) = \frac{1}{|Y_{ij}|} \sum_{y \in Y_{ij}} \mathbf{w}_y^\top \mathbf{d}(x, y) + b \,, \tag{2}$$

where $\mathbf{W}$ is a set of weight vectors $\mathbf{w}_y$ for all $y \in Y_{ij}$. This significantly more expressive distance function does not suffer from the problems that plague *nearest instance*. Note that we have added a bias term, b, to the instance based distance function to allow negative values. The weight vector $\mathbf{w_y}$ is D-dimensional, where D is the number of different features extracted for each view. Note also that each example view has a different weight vector. Due to this, the functions do not define a true distance metric, as they are asymmetric. This is advantageous since different examples may have different sets of features that are better for distinguishing them from other examples, or views.

When learning the weight vector for an instance, it is necessary to distinguish between category and instance classification. For *instance recognition*, the weight $\mathbf{W}_{ij}$ defining the distance function for the j-th instance in category i can be learned using the following L_2 regularized loss function:

$$\sum_{x \in Y_{ij}} L(-f(x, \mathbf{W}_{ij})) + \sum_{x \in Y \setminus Y_{ij}} L(f(x, \mathbf{W}_{ij})) + \lambda ||\mathbf{W}_{ij}||_2, \quad \text{subject to } \mathbf{W}_{ij} \succeq 0, \tag{3}$$

where we have chosen $L(z) = \max(0, 1 - z)^2$, the squared hinge loss. The non-negative weight constraint ensures that large feature differences can never imply similarity. The first term penalizes misclassification of views $x \in Y_{ij}$ that belong to the same instance. The second term similarly penalizes misclassification of negative examples, or views, by incurring a loss when their distance is small. Note that the negative examples also include views of different instances that belong to the same category i. The final term serves standard L_2

parameter regularization. This objective function is convex and is easily optimized using standard optimization algorithms.

For *category recognition*, we learn the instance-based distance by minimizing the following L_2 regularized loss:

$$\sum_{x \in Y_i} L(-f(x, \mathbf{W}_{ij})) + \sum_{x \in Y \setminus Y_i} L(f(x, \mathbf{W}_{ij})) + \lambda ||\mathbf{W}_{ij}||_2, \quad \text{subject to } \mathbf{W}_{ij} \succeq 0, \quad (4)$$

where $Y_i = \bigcup_{s=1}^{N_i} Y_{is}$ and N_i is the number of instances in the $i-th$ category. The key difference between the instance recognition and the category recognition loss is that in the former, only the views of the same instance are positive examples, whereas in the latter the views of *all* instances in the same category become positive examples.

III. Example Selection via Group-Lasso

An important property of the *instance based distance* that we defined in Section II is that it allows for data sparsification. This is achieved by replacing L_2 regularization in (3) with Group-Lasso [23, 15], resulting in the following objective function:

$$\sum_{x \in Y_{ij}} L(-f(x, \mathbf{W}_{ij})) + \sum_{x \in Y \setminus Y_{ij}} L(f(x, \mathbf{W}_{ij})) + \lambda \sum_{y \in Y_{ij}} ||\mathbf{w}_y||_2 \quad \text{subject to } \mathbf{W} \succeq 0 \quad (5)$$

Here, the first two terms optimize over individual components of the instance weight vector, and the third, Group-Lasso, term drives the weight vectors of individual views toward zero. Group-Lasso achieves this by grouping the weight components of individual views in the penalty term. In contrast to previous work that make use of the Group-Lasso for encouraging feature sparsity, here we use Group-Lasso to encourage data sparsity. In other words, optimizing this objective function gives a supervised technique for choosing a subset of representative examples, or views. If the Group-Lasso drives an entire weight vector $\mathbf{w}_y$ to $\mathbf{0}$, the corresponding example no longer affects the decision boundary and has effectively been removed by the optimization. The degree of sparsity can be tuned by varying the λ parameter. Intuitively, data sparsity is often possible because many examples may lie well within the decision region or are densely packed together. Removing such examples would reduce the magnitude of the regularization term while having little or not effect on the loss terms. Each data point is only one of many that contribute to the *instance-based distance* and redundant examples would not significantly influence the decision boundary.

The advantage of data sparsification using the proposed objective function is twofold. As explained above, the proposed technique can remove redundant and uninformative examples. Secondly, removing examples from consideration at test time results in computational cost savings which counteracts the data-size-dependent time complexity of nearest neighbor techniques. For category level, the group lasso based instance distance learning uses the following objective function

$$\sum_{x \in Y_i} L(-f(x, \mathbf{W}_{ij})) + \sum_{x \in Y \setminus Y_i} L(f(x, \mathbf{W}_{ij})) + \lambda \sum_{y \in Y_{ij}} ||\mathbf{w_y}||_2 \quad \text{subject to } \mathbf{W} \succeq 0 \quad (6)$$

IV. Experiments

We apply the proposed subclass distance function learning to two related object recognition tasks: category recognition and instance recognition. In category recognition, the system is trained on several objects belonging to each category and the task is to classify a never-before-seen object into one of the categories. In the instance recognition task, the system is presented with multiple views of each object, and the task is to classify never-before-seen views of these same objects. The experimental results in this section demonstrate that our technique obtains good performance on both recognition tasks, particularly when taking full advantage of both shape and visual information available from the sensor. The technique is able to not only automatically sparsify training data, but it also exceeds the performance of several alternative approaches and baselines, even after sparsification.

A. RGB-D Data Set

We evaluate our technique on a novel data set consisting of images of objects spun around on a turntable (the *RGB-D data set*). The data set consists of 159 object instances in 31 categories. Table I gives a breakdown of the number of instances in each category. The images are collected with the RGB-D camera that can simultaneously record both color image and depth data at 640×480 resolution. In other words, each 'pixel' in the RGB-D frame contains four channels: red, green, blue and depth. The location of each pixel in 3D euclidean space can be computed using known sensor parameters, meaning that in essence each RGB-D pixel is a 6-dimensional point. Each object was placed on the turntable and rotated. Data was recorded from three viewing angles, at approximately 30, 45 and 60 degrees with the horizon. We used around 33 views at each viewing angle, giving around 100 views per instance, or 15900 RGB + Depth images in total, each of which serves as a data point in training or testing. Fig. 2 shows some example views from the data set.

B. Segmentation and Feature Extraction

Since we know that the object lies on a table, which is a flat surface, we can segment out the object by performing RANSAC plane fitting [7] to find the table plane and then remove it. Next, we extract features from each view. The presence of synchronized visual and 3D data greatly enhances the amount of information available for performing object recognition and our technique naturally combines multiple features into a single framework. We first extract a set of features capturing the shape of a view. For each RGB-D point, we compute spin image signatures [11], which capture the local shape information around the point. To incorporate spatial information, we partition the points into a $3 \times 3 \times 3$ spatial grid. To generate a final fixed length feature vector

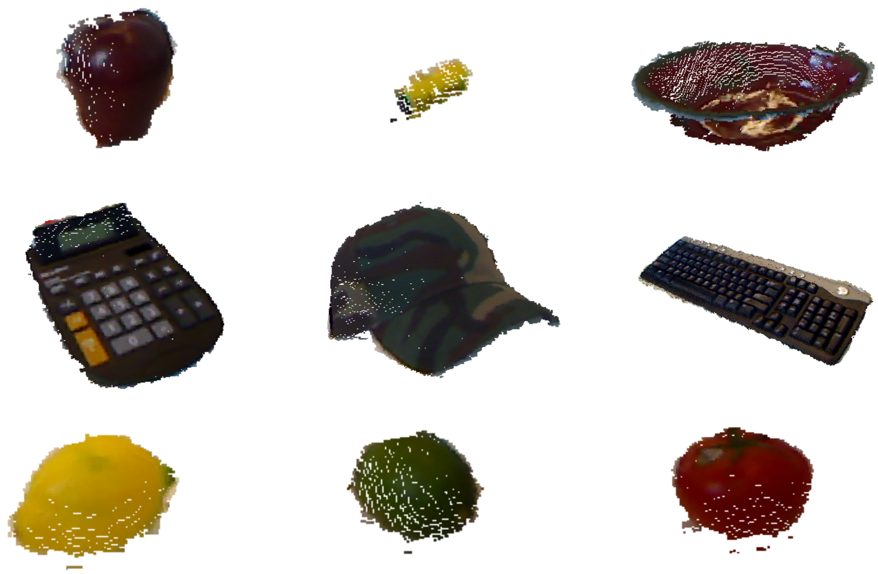

Fig. 2. Views of objects from the RGBD data set. From left to right, top to bottom, they are apple, battery, bowl, calculator, cap, keyboard, lemon, lime, tomato.

Category	apple	battery	binder	bowl	calculator	camera	cap
# of Instances	5	6	3	4	5	3	4
Category	cell phone	cereal box	comb	crayon	flashlight	keyboard	lemon
# of Instances	5	5	5	7	5	5	6
Category	lime	lock	mug	orange	pear	pepper	pitcher
# of Instances	4	4	6	6	6	3	7
Category	plate	pliers	potato	scissor	soda can	sponge	stapler
# of Instances	7	6	6	5	6	4	6
Category	tomato	toothbrush	water bottle				
# of Instances	5	4	6				

TABLE I

THE CATEGORY AND INSTANCE COMPOSITION OF THE RGBD DATA SET.

from this set of local descriptors, we generate efficient match kernel (EMK) features using random fourier sets as proposed in [2]. Each grid cell has a 100-dimensional EMK feature and so overall we obtain 27 EMK spin image descriptors each of length 100. We also include as shape features the width, depth and height of the segment's bounding box. This gives us a total of 30 shape descriptors.

To capture the visual appearance of the view, we extract densely-sampled SIFT [13] features and generate EMK features (of fixed length 1000) with them using the same technique described above for spin image signatures. We also extract other visual features including color histograms and texton histograms for a total of 42 features.

Evaluation

The learning algorithms that we evaluated are:

- IDL: Our proposed instance-based distance learning algorithm with L_2 regularization.
- IDL SPARSE: Sparse instance-based distance learning learning with Group-Lasso regularization.
- NIC: The nearest instance classifier baseline described in Section II.
- EB LOCAL: Exemplar-based local distance function learning technique first proposed in [14].

Algorithm	Accuracy (# / % training data retained)
IDL	84.0 (12554/100.0)
IDL SPARSE	86.5 (11572/92.1)
	86.1 (9904/78.9)
	86.1 (5079/40.5)
	85.8 (3418/27.2)
	85.2 (2384/19.0)
	84.6 (2085/16.6)
NIC	51.5 (12554/100.0)
EB Local	80.0 (12554/100.0)

TABLE II

CLASSIFICATION PERFORMANCE OF OUR TECHNIQUES AND ALTERNATIVE APPROACHES ON THE RGB-D DATA SET. IDL IS INSTANCE BASED DISTANCE LEARNING, IDL SPARSE IS INSTANCE BASED DISTANCE LEARNING WITH DATA SPARSITY, NIC IS NEAREST INSTANCE CLASSIFIER, AND EB LOCAL IS EXEMPLAR-BASED LOCAL DISTANCE LEARNING.

For category recognition, we randomly select one instance per category for as test data, while training on the rest, for 5 train/test splits. For instance recognition, we randomly select 80% of the views for training and test on the remaining 20%, once again repeated for 5 train/test splits.

Features	Classification Accuracy	
	Category Recognition	Instance Recognition
Shape	71.6	68.6
Vision	73.6	85.2
Shape+Vision	84.0	93.0

TABLE III
CLASSIFICATION PERFORMANCE OF THE *instance-based distance learning* TECHNIQUE WITH DIFFERENT FEATURES.

Table II shows the overall classification accuracies of the different algorithms on the category recognition task. As can be seen from the results, our technique significantly outperforms the baseline *nearest mean distance* and substantially improves upon the performance of a competitive exemplar-based local distance method. IDL SPARSE is able to sparsify the data considerably (by almost a factor of $\frac{1}{10}$) without causing any significant loss in accuracy. The fact that IDL SPARSE actually slightly outperforms IDL despite removing so many data points is particularly encouraging. A histogram of the number of examples retained for a subset of objects (one per category) after IDL SPARSE optimization is shown in Fig. 4. From this, it can be seen that the technique retains a different number of examples of each object, trading off the classification loss during training with the desired amount of sparsity controlled by the λ parameter in front of the Group-lasso term.

Fig. 3 shows the confusion matrices between the 31 categories for the (top) the instance-based distance learning technique (IDL) and (bottom) the instance-based distance learning technique with data sparsification (IDL SPARSE).

In any recognition task it is important to extract the correct set of features. However, the appropriate set of features may be different for the two tasks. For example, if we wish to distinguish between soda cans and keyboards, both shape and visual information is very important. On the other hand, if we are trying to distinguish between a Coke can and a 7up can, the shape is completely uninformative. As described in Section II, it is straightforward to apply IDL to the case of multiple features. To verify that our technique is indeed able to take advantage of both shape and visual information available from the RGB-D camera, we additionally performed feature ablation experiments. We evaluated the performance of the IDL technique with only shape-based features and with only visual-based features. Table III shows the resulting classification accuracies for both the category and instance task. The result of using both shape and visual features from Table II is repeated here for ease of comparison. Consistent with our intuitions, the relative importance of shape-based and vision-based features differs for the two tasks, with visual appearance being much more important for instance recognition. The fact that combining both shape and visual features enables our technique to perform better on both tasks demonstrates that our technique provides a common framework for solving both tasks.

V. DISCUSSIONS

In this work we have studied joint object category and instance recognition using a large RGB-D (color+depth) dataset of everyday objects. Our work are of interest both in terms of algorithm design and of the empirical validations on appearance and depth cues for recognition. Our key insight is that because an object category consists of object instances, there is a natural division of a category into subclasses, and this motivates our use of the per-instance instance. We show that by joining learning the weights in the per-instance distance, we outperform local distance learning methods. The use of Group-Lasso allows us to find a compact representation of each object instance as a small set of views, and this sparsification further improves recognition accuracy.

We have empirically validated our models for various choices of objective functions and the use of vision (RGB) and shape (depth) cues. We show that jointly shape+vision achieves much higher performance than either set of cues alone, for both category and instance recognition. Considering the fast advances of RGB-D camera hardware, these results are of great interest and send a clear message: that the combination of RGB and depth will find many uses in object recognition and other perception tasks. We also confirm that vision (RGB) cues are more useful for instance recognition, where texture plays a central role, while performing about the same as shape cues on category recognition.

Furthermore, with the ever increasing size of data sets available on the world wide web, *sparsification* of such data will become a more and more important issue. In contrast to online, greedy selection algorithms, we here introduce Group-Lasso as an alternative sparsification approach. While the current technique assumes an offline setting, the development of online Group-Lasso style sparsification is an interesting and promising direction for future work.

REFERENCES

[1] F.R. Bach, G.R.G. Lanckriet, and M.I. Jordan. Multiple kernel learning, conic duality, and the SMO algorithm. In *Proceedings of the twenty-first international conference on Machine learning*, page 6. ACM, 2004.

[2] L. Bo and C. Sminchisescu. Efficient Match Kernel between Sets of Features for Visual Recognition. In *Advances in Neural Information Processing Systems*, December 2009.

[3] Canesta. `http://www.canesta.com/`.

[4] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-fei. ImageNet: A Large-Scale Hierarchical Image Database. In *Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[5] Sergio Escalera, David M. J. Tax, Oriol Pujol, Petia Radeva, and Robert P. W. Duin. Subclass problem-dependent design for error-correcting output codes. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(6):1041–1054, 2008.

[6] L. Fei-Fei, R. Fergus, and P. Perona. One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 28(4):594–611, 2006.

[7] Martin A. Fischler and Robert C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395, 1981.

[8] A. Frome, Y. Singer, and J. Malik. Image retrieval and classification using local distance functions. In *Advances in Neural Information Processing Systems (NIPS)*, 2006.

[9] A. Frome, Y. Singer, F. Sha, and J. Malik. Learning globally-consistent local distance functions for shape-based image retrieval and classication. In *Proc. of the International Conference on Computer Vision (ICCV)*, 2007.

[10] A. Globerson and S. Roweis. Metric learning by collapsing classes. *Advances in Neural Information Processing Systems*, 18:451, 2006.

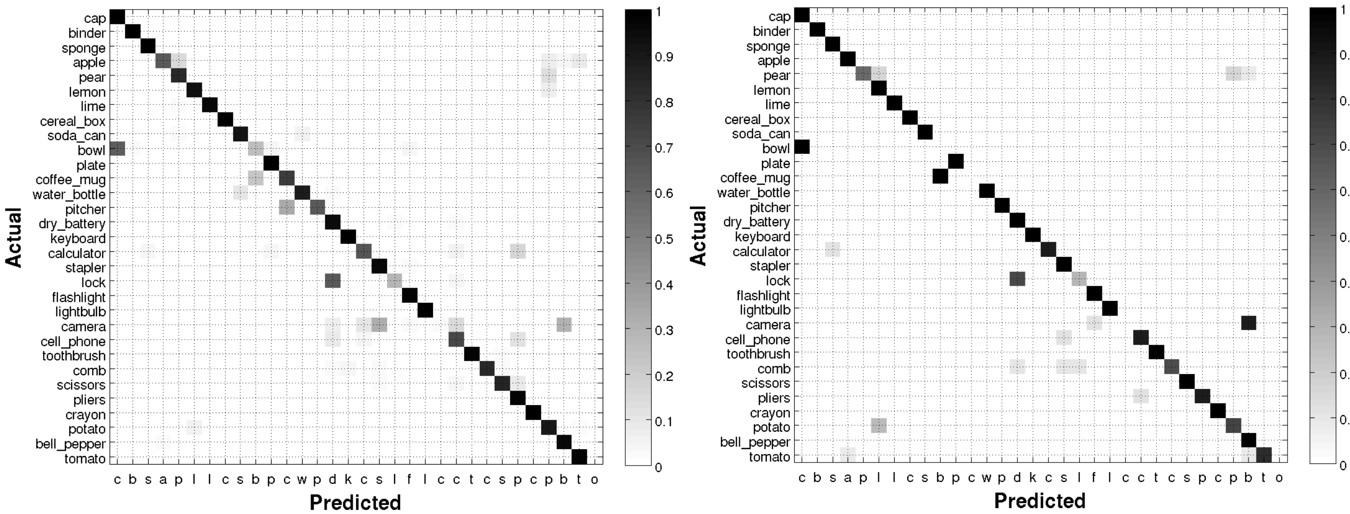

Fig. 3. Confusion matrices (row-normalized) between the 31 categories for (left) the *instance-based distance learning* technique and (right) the *sparse instance-based distance learning*.

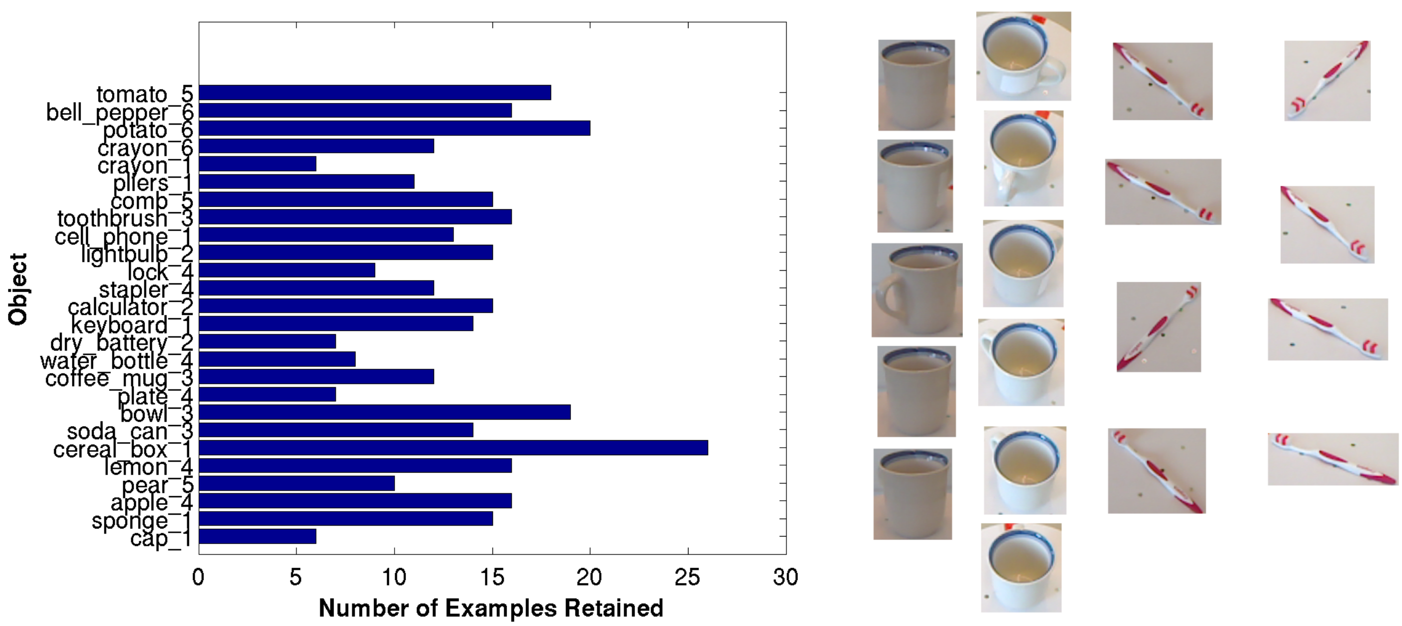

Fig. 4. Data selection with Group-Lasso: (Left) histogram of the number of views chosen for a subset of objects and (right) all the views that were chosen for a coffee mug and a toothbrush.

[11] A. Johnson and M. Hebert. Using spin images for efficient object recognition in cluttered 3D scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 21(5), 1999.

[12] K. Lai and D. Fox. Object Recognition in 3D Point Clouds Using Web Data and Domain Adaptation. *The International Journal of Robotics Research*, 2010.

[13] David G. Lowe. Object recognition from local scale-invariant features. *Computer Vision, IEEE International Conference on*, 2:1150, 1999.

[14] T. Malisiewicz and A. Efros. Recognition by association via learning per-examplar distances. In *Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[15] Lukas Meier, Sara Van De Geer, Peter Bhlmann, and Eidgenssische Technische Hochschule Zrich. The group lasso for logistic regression. *Journal of the Royal Statistical Society, Series B*, 2008.

[16] PrimeSense. `http://www.primesense.com/`.

[17] D. Ramanan and S. Baker. Local Distance Functions: A Taxonomy, New Algorithms, and an Evaluation. In *Proc. of the International Conference on Computer Vision (ICCV)*, 2009.

[18] M. Schultz and T. Joachims. Learning a distance metric from relative comparisons. In *Advances in Neural Information Processing Systems (NIPS)*, page 41. The MIT Press, 2003.

[19] N. Shental, T. Hertz, D. Weinshall, and M. Pavel. Adjustment learning and relevant component analysis. pages 181–185. Springer, 2002.

[20] Cor J. Veenman and Marcel J. T. Reinders. The nearest subclass classifier: A compromise between the nearest mean and nearest neighbor classifier. *IEEE Trans on Pattern Analysis and Machine Intelligence*, 27:1417–1429, 2005.

[21] K.Q. Weinberger and L.K. Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research (JMLR)*, 10:207–244, 2009.

[22] E.P. Xing, A.Y. Ng, M.I. Jordan, and S. Russell. Distance metric learning with application to clustering with side-information. *Advances in Neural Information Processing Systems (NIPS)*, pages 521–528, 2003.

[23] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *JOURNAL-ROYAL STATISTICAL SOCIETY SERIES B STATISTICAL METHODOLOGY*, 68(1):49, 2006.