LCAV-31: A Dataset for Light Field Object Recognition

Alireza Ghasemi , Nelly Afonso and Martin Vetterli

AudioVisual Communications Laboratory
École Polytechnique Fédérale de Lausanne

ABSTRACT

We present LCAV-31, a multi-view object recognition dataset designed specifically for benchmarking light field image analysis tasks. The principal distinctive factor of LCAV-31 compared to similar datasets is its design goals and availability of novel visual information for more accurate recognition (i.e. light field information).

The dataset is composed of 31 object categories captured from ordinary household objects. We captured the color and light field images using the recently popularized Lytro consumer camera. Different views of each object have been provided as well as various poses and illumination conditions. We explain all the details of different capture parameters and acquisition procedure so that one can easily study the effect of different factors on the performance of algorithms executed on LCAV-31.

Moreover, we apply a set of basic object recognition algorithms on LCAV-31. The results of these experiments can be used as a baseline for further development of novel algorithms.

Keywords: Plenoptic Function, Light Field Imaging, Object Recognition, Computer Vision Benchmarking

1. INTRODUCTION

Light field imaging has received increased attention in recent years mostly due to availability of consumer light field cameras such as Lytro[8] and Raytrix *. In this emerging imaging technology rays from multiple views are captured from a scene. The improved sensing capability and the extra information captured by a light field camera enable novel lines of research in order to significantly improve current vision tasks such as object recognition, reconstruction and depth mapping.

Many light field processing algorithms have been proposed in recent years due to the increased attention toward this area. The need to evaluate different aspects of these algorithms has led to various light field image sets proposed so far. The structure and associated ground truth information of any of these datasets have been optimized for a specific task such as reconstruction or depth estimation.

On the other side, many object recognition benchmarks are already known among the computer vision community. There is a huge diversity regarding the objectives, scope, capturing setup and other properties between these datasets. However, they mostly consist of traditional color or grayscale images. This restricts their use in evaluation of algorithms which utilize data of modern sensors such as depth or light fields. This issue and other drawbacks of current object recognition datasets such as bias and lack of object diversity bolds the need to construct a novel dataset with more information available per object than traditional color images.

The goal of this project is to build a dataset using plenoptic informations with a purpose of object recognition. Plenoptic photography captures the available light in a scene coming from more then one direction. It works by breaking up the main image with an array of microlenses over an image sensor. The camera software then uses this data to determine the general directions of incoming light rays. Such a dataset would be the first in its category. Indeed, most of datasets for object recognition uses traditional images. There exists some works on plenoptic images but they don't deal with object recognition. To start the project, we have to answer this question : How to build a good object recognition dataset ?

LCAV-31 is the first light field object recognition dataset which provides a unified framework for benchmarking light field retrieval, object recognitions and tracking systems. It provides the possibility to evaluate effectiveness of different sensed information such as color and light field in object recognition accuracy.

*http://www.raytrix.de

Computational Imaging XII, edited by Charles A. Bouman, Ken D. Sauer, Proc. of SPIE-IS&T Electronic Imaging,
SPIE Vol. 9020, 902014 · © 2014 SPIE-IS&T · CCC code: 0277-786X/14/$18 · doi: 10.1117/12.2041097

We first study important properties of different state-of-the-art datasets. We analyze various types of bias in dataset creation and their effect on evaluation results. We also analyze the diversity and robustness of current datasets. Then we explain how we avoid the biases and achieve a high level of diversity in LCAV-31.

Finally, we also apply a set of basic object recognition algorithms on LCAV-31. The results of these experiments can be used as a baseline for further development of novel algorithms.

2. EXISTING DATASETS

There are a lot of datasets which are currently used by the computer vision community to benchmark novel object recognition algorithms. In this section we review a number of most important ones among them.

Two of the famous datasets in image retrieval are Caltech-101[6] and Caltech-256.[10] They are huge sets of different object categories (9144 images with 102 categories for Caltech 101 and 30607 images with 257 categories for Caltech 256). Caltech has, therefore, become a de facto standard for evaluating algorithms for multi-class category-level recognition.

These datasets show diverse types of images with different poses, illuminations, sizes, camera setup and background texture. Often objects are dominant in the images with small or medium background clutter and there is usually only one object per image. There are also drawings in the datasets.

Caltech-256 has more images per category than Caltech 101 (minimum number per category is increased from 31 to 80). Moreover, in Caltech 101, object categories were consistently right-left aligned whereas in Caltech 256, they were not.[11]

The UIUC Car Detection Dataset[2] is another widely used image database whose images show side views of cars. the dataset is decomposed into different categories with and without variations in scale of the cars. The images were acquired from both still and dynamic scenes. They include different variations in background complexity and the amount of occlusion.

In the MIT CSAIL LabelMe project,[3] the goal is to construct a huge set of images of natural scenes together with manual segmentations and annotations.

LabelMe contains indoor and outdoor objects in office and urban environments. The database provides annotations for more than 30 objects in context. Some objects provide additional information like point of view. Images are also labeled according to scene information. Some of the frames provide specific place information. Images come from different sources (webcam, digital cameras, and images from the web). The images are high resolution and cover a wide field of view, providing rich contextual information.

However,the most important problem with LabelMe is that the annotations are sparse and are not distributed evenly in all frames.

The PASCAL Visual Object Classes (VOC) challenge[5] is an annual competition and workshop in object classification and detection which provides a dataset of images and annotations. The images are annotated consumer photographs with a lot of variations in background, cluttering and other aspects.

In[12] the SUN12 dataset is introduced. This dataset is a collection of annotated images covering a large variety of environmental scenes, places and the objects within. Currently, there are 131072 Images, 908 Scene categories, 249522 Segmented objects and 3819 Object categories.

The GRAZ dataset was introduced in.[14] The goal of the dataset is addressing the need of images having many different instances of the object category at different times and scales, viewed from various angles and with high background clutter. It contains samples of classes *persons*, *bikes* and *none*.

Based on localization results which reveal that certain methods are biased toward scenes with many background features, a second database GRAZ-02, has been introduced.

the authors in[14] claim that on the Caltech and UIUC datasets, only 25% to 50% of homogeneity regions are located on the object. The remaining hypotheses actually focus on contextual (background) information not the objects themselves.

Therefore, GRAZ-02 has been carefully balanced with respect to background, so that similar backgrounds occur for all categories. Furthermore, the complexity of the object appearances was increased and a third category of images was added.

ImageNet[1] is an image dataset organized according to the WordNet hierarchy. The WordNet[†] is a large lexical database of the English language. In ImageNet, images of each concept are manually controlled for quality and annotated.

Another widely-known dataset is that of the Microsoft Research at Cambridge.[7] The objective of the creators of this dataset is to automatically generate a large number of images for a specified object class. A multi-modal approach employing both text, meta data and visual features is used to gather many high-quality images from the web. Candidate images are obtained by a text based web search querying on the object identifier (e.g. the word penguin). The web pages and the images they contain are downloaded. The task is then to remove irrelevant images and re-rank the remainder. The images are re-ranked based on the text surrounding the image and meta data features.

The version 2 of this dataset is an extension of the version 1 and contains 591 images with 23 object classes and accurate pixel-wise labeled images. Though it contains 23 object classes, only 21 classes are commonly used. The unused labels are (void, horse, mountain) due to background or too few training samples. The dataset is commonly used for full scene segmentation, and may also be used for object instance segmentation, as the current annotation also contains individual object instances next to pure class annotation.

The RGB-D Object Dataset[13] is a large dataset of 300 common household objects. The objects are organized into 51 categories arranged using WordNet hypernym-hyponym relationships (similar to ImageNet). These categories are a subset of the categories in ImageNet. This dataset was recorded using a Kinect style 3D camera that records synchronized and aligned 640x480 RGB and depth images at 30 Hz. This is the only one presented in this report that doesn't use "traditional" images. Each object was placed on a turntable and video sequences were captured for one whole rotation. For each object, there are 3 video sequences, each recorded with the camera mounted at a different height so that the object is viewed from different angles with the horizon.

2.0.1 Columbia Object Image Library

There are two main datasets called Columbia Object Image Library : COIL-20 and COIL-100.[15] Images in these datasets are image of small objects like cup, box of medicine or figurine. The objects were placed on a motorized turntable with a black background. The turntable was rotated through 360 degrees to vary object pose with respect to a fixed color camera. Images of the objects were taken at pose intervals of 5 degrees. The data is therefore images of objects with 72 different poses (there are 72 images per object). Most objects appeared in the center of the images.

COIL-20 has two sets of images. The first set contains 720 unprocessed images of 10 objects. The second contains 1440 size normalized images of 20 objects.

COIL-100 is a database of 7200 color images of 100 different objects. The images were sized normalize.[15]

3. THE PROPOSED DATASET

3.1 Specifications

Having studied the benefits and drawbacks of current object recognition datasets, we will introduce in this section our dataset LCAV-31 which tries to aggregate the best practices in dataset design and avoid biases and drawbacks.

We chose household and office objects to have a dataset with popular objects that we can find in a lot of different scenes.

We also wanted to build a dataset with different levels of difficulty. So, we picked different objects with different difficulties. For example, little objects or transparency can be hard to recognize as you can see on figure 1. Indeed, the cup is transparent and in some part, we can mix it up with the background.

[†]http://wordnet.princeton.edu/

Regarding the environmental settings we captured pictures of each object well illuminated with a monochrome background for the easy part (Figure 2) and with daylight for moderate difficulty (Figure 6).

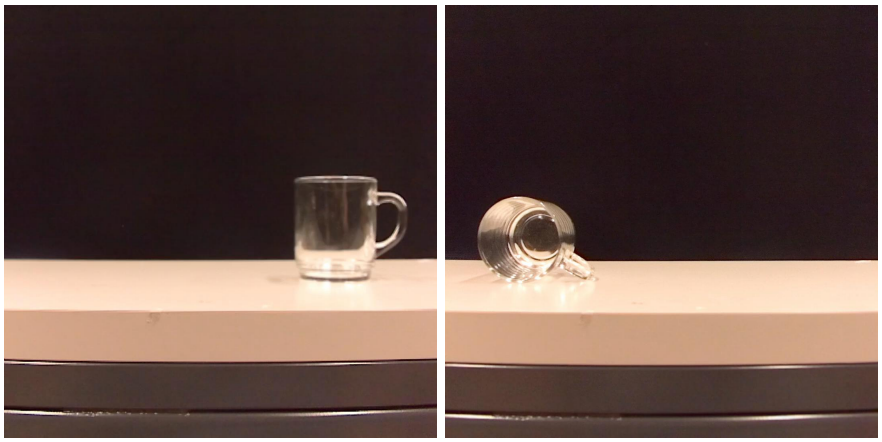

Figure 1. example of photographs with "difficult" objects (transparent)

3.1.1 Avoidance of Bias

To build LCAV-31, we wanted to avoid the biases that we mentionsed in the previous section.

- **The Label Bias** : To try to avoid this bias, we can use WordNet hyponym/hypernym relations. Therefore, as this set of words is known and not misused, the labels should be the same in a lot of datasets. Moreover, in our case, the problem about "what to label" is not really relevant as we shoot objects that we will define before. The only problem can be about what sort of class we choose. It means, for example, if we choose the class Phone or Office phone and Mobile phone. We have to choose our level of precision between a general class and its subclasses. In this dataset, we choose to be as precise as possible.
- **The Capture Bias** : We will take pictures of the same object under different angles (different heights relative to the floor and several angles during a revolution of each object at each height.). We will also take photographs with the object in different part of the photography without forget the center. Like that, we will try to have all the possible poses and we will not forget the typical ones either. You can see different poses for the classes Router, Teddy bear and Stapler in figure **??**.

3.2 The Dataset

For the moment, there are 31 classes of objects caught with a Lytro camera.

The list of classes whose names are referenced in WordNet are :

Cup, Punch pliers, Picture frame, Scissors, Sunglasses, Ruler, Eraser, Pen, Book, Lighter, Mouse, Sponge, Adhesive tape, Plate, Bulb, Stapler, Keyboard, Office phone, Hand tool, Wool, Vise, Meter, Flowerpot, Clock, Teddy bear, Spinning top, Router, Remote, Food processor, Thermos bottle, Bowl.

Some of the classes of the dataset have several iterations. We caught on camera 3 sunglasses, 2 scissors, 4 pens, 2 books, 4 mouses, 2 plates, 3 staplers, 3 keyboards, 2 office phones, 3 wools, 2 meters, 4 cups and 4 adhesive tapes.

Figure 2 depicts some of the objects in LCAV-31.

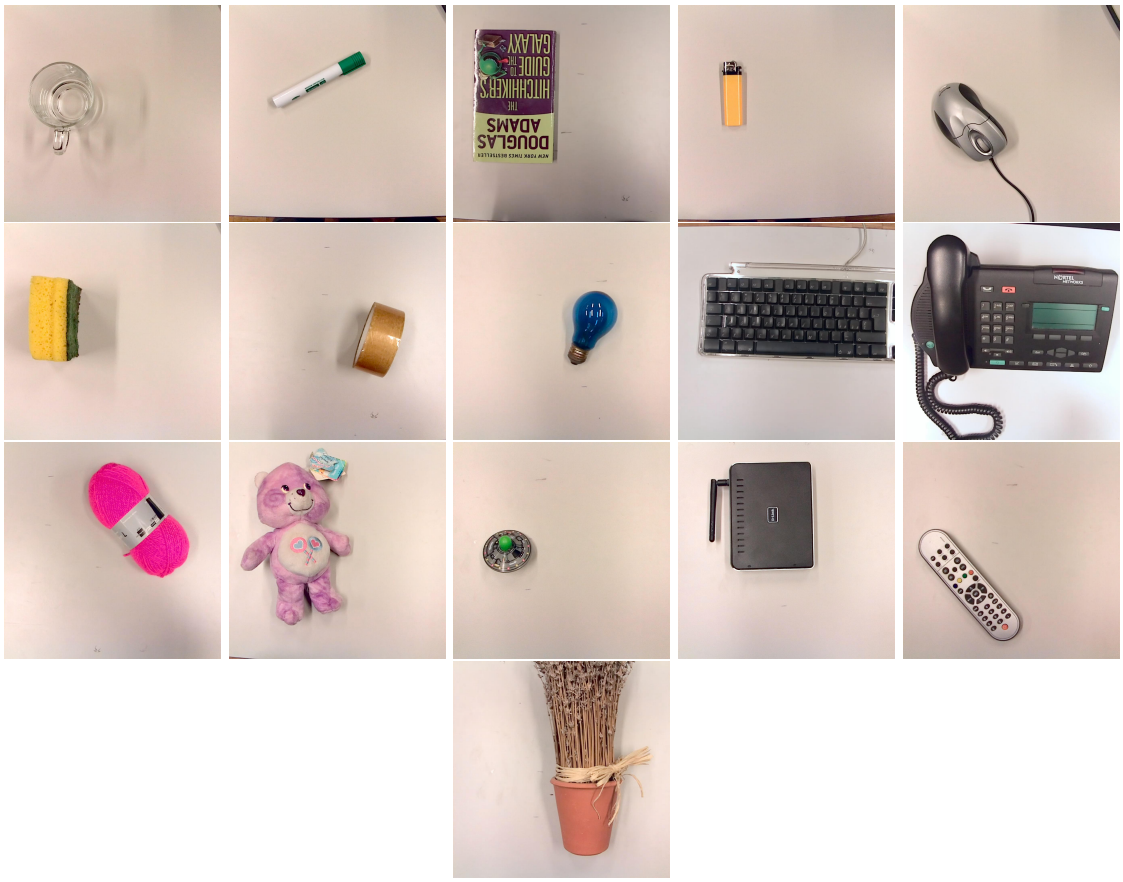

Figure 2. Examples of objects in our dataset (no background clutter, taken in a dark room). Categories : Cup, Pen, Book, Lighter, Mouse, Sponge, Adhesive tape, Bulb, Keyboard, Office phone, Wool, Teddy bear, Spinning top, Router, Remote and Flower pot.

Currently, the main part of the dataset consists of photographs of one object taken in a dark room with good illumination and a background without clutter. Objects are placed on a table with a height of 63 centimeters.

In this darkroom, there are three different placement of camera.

- The first one is above the objects at about 44 centimeters from the table (Figure 3).
- The second one is next to the objects with a distance of 1 meter (Figure 4).
- The third one is next to the objects with a distance of 0,5 meter (Figure 5).

For the second and the third ones, the height of the camera from the floor is about 65 centimeters.

There are also photographs with the same environmental setup but with several objects per image. The camera is next to the objects with a distance of 0,5 meter and a height from the floor of about 65 centimeters. The classes present in these part of the dataset are Cup, Wool, Adhesive tape and Plate.

Finally, a little part of the dataset consists of photographs of the class Clock taken under daylight and with a background without clutter. The camera is placed next to the object on the same table and with a distance of about 0,5 meter. You can see the photographs on figure 6.

Figure 3. example of photographs with the camera above the objects. Categories : plate, stapler, hand tool and thermos bottle

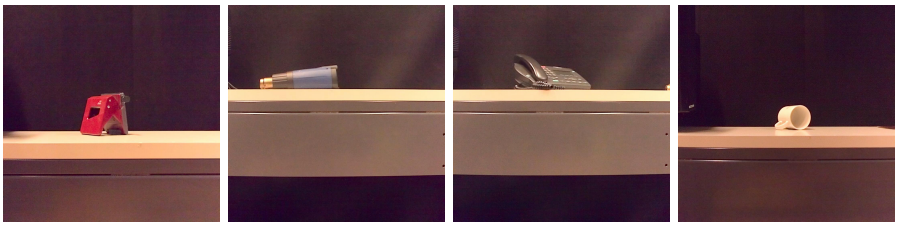

Figure 4. example of photographs with the camera next to objects and a distance of one meter. Categories : punch pliers, hand tool, phone and cup

4. PRELIMINARY EXPERIMENTS

In this section we explain some initial experiments done on the datasets that can be used for further benchmarking tasks.

We compared a light-field specific method described in[9] called STILT with DenseSIFT, which is a recent version of the SIFT descriptor claiming to be faster than it.[16] We also implemented and tested the Histogram of Oriented Gradients (HOG) approach[4] which has proven successful for outdoor classification tasks.

For the classification algorithm, we used a simple nearest-neighbor approach as well as more complicated machine learning methods so that we can see the effects of the extracted features and the utilized recognition approach separately . We also utilized different distance measures to see the effects.

Table 1 depicts the results of STILT, HOG and DenseSIFT. The initial results show that the dataset is a difficult one and therefore the results can be highly improved by employing more advanced approaches. The poor results of DenseSIFt and HOG are partly because of the low spatial resolution of Lytro cameras which prevents significant features of the object to be precisely detect. This makes a challenge for recognition algorithms that can invest on this dataset.

	STILT	*DenseSIFT*	*HOG*
Average Accuracy with Manhattan Distance and Nearest Neighbor	**65**	58	57
Average Accuracy with Euclidean Distance and Nearest Neighbor	**64**	59	55
Average Accuracy with Tanimoto Distance and Nearest Neighbor	**64**	55	58
Average Accuracy with Manhattan Distance and Logistic Regression	**67**	53	70
Average Accuracy with Euclidean Distance and Logistic Regression	**67**	62	68
Average Accuracy with Tanimoto Distance and Logistic Regression	**67**	60	61
Average Accuracy with Manhattan Distance and Gaussian Process	**67**	59	72
Average Accuracy with Euclidean Distance and Gaussian Process	**62**	64	68
Average Accuracy with Tanimoto Distance and Gaussian Process	**64**	61	64

Table 1. Results of Running Different Algorithms on the LCAV-31 Dataset in the Binary Classification Setup. Accuracy is the Percentage of Correct Classifications

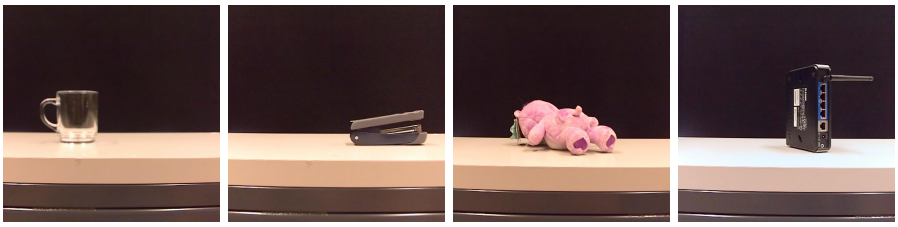

Figure 5. example of photographs with the camera next to objects and a distance of half a meter. Categories : cup, stapler, teddy bear and router

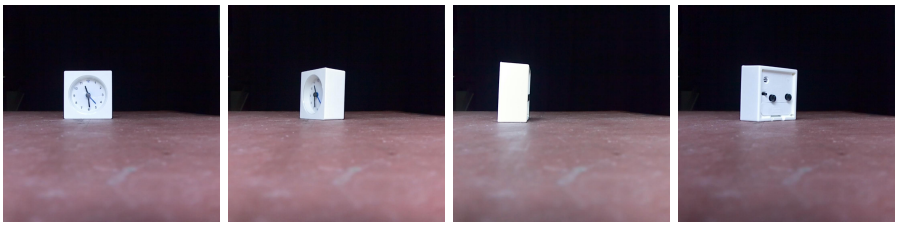

Figure 6. photographs of Clock under daylight

SUMMARY

The increased popularity of consumer light field cameras in recent years and the attention toward light field imaging within the computer vision research community strengthens the hypothesis that in the near future most of the mainstream digital photography will be based on light field imaging. The extra information contents of light field images can be very efficiently used for improving many of the current computer vision tasks, especially object recognition and image retrieval algorithms. Utilizing the vast amount of information inherent in light field images requires adaptation of current computer vision algorithms or developing completely new algorithm from scratch.

Availability of a precisely designed task-specific dataset is a core requirement for global benchmarking and evaluation of such novel and adapted computer vision algorithms. There are already massively adopted image sets for tasks such as light-field depth estimation and segmentation. In this paper we propose a precisely designed dataset for light-field object recognition. The dataset has been constructed under controlled settings which facilitates reasoning and justification about the results obtained it. A consumer-available Lytro camera has been used to capture the light field images

ACKNOWLEDGMENTS

This work has been co-funded by the Committee for Technological Innovations (CTI) in Switzerland.

	STILT	*DenseSIFT*	*HOG*
Average Accuracy with Manhattan Distance and Nearest Neighbor	**53**	41	44
Average Accuracy with Euclidean Distance and Nearest Neighbor	**50**	41	45
Average Accuracy with Tanimoto Distance and Nearest Neighbor	**50**	43	45
Average Accuracy with Manhattan Distance and Logistic Regression	**56**	44	41
Average Accuracy with Euclidean Distance and Logistic Regression	**56**	47	39
Average Accuracy with Tanimoto Distance and Logistic Regression	**54**	49	39
Average Accuracy with Manhattan Distance and Gaussian Process	**67**	44	42
Average Accuracy with Euclidean Distance and Gaussian Process	**64**	44	38
Average Accuracy with Tanimoto Distance and Gaussian Process	**61**	41	34

Table 2. Results of Running Different Algorithms on the LCAV-31 Dataset in the Multi-Class Classification Setup. Accuracy is the Percentage of Correct Classifications. All Pairs of Classes have been tested.

REFERENCES

[1] Imagenet. "http://www.image-net.org", 2010.

[2] Agarwal, S., Awan, A., and Roth, D. Learning to detect objects in images via a sparse, part-based representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence 20*, 11 (2004), 1475–1490.

[3] B. C. Russell, A. Torralba, K. P. M. W. T. F. Labelme: a database and web-based tool for image annotation. *International Journal of Computer Vision 77*, 1-3 (May 2008), 157–173.

[4] Dalal, N., and Triggs, B. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on* (2005), vol. 1, IEEE, pp. 886–893.

[5] Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., and Zisserman, A. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.

[6] Fei-Fei, L., Fergus, R., and Perona, P. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Computer Vision and Image Understanding 106*, 1 (2007), 59–70.

[7] Florian Schroff, A. C., and Zisserman, A. Harvesting image databases from the web. *IEEE Transactions On Pattern Analysis and Machine Intelligence 33*, 4 (April 2011), 754–766.

[8] Georgiev, T., Yu, Z., Lumsdaine, A., and Goma, S. Lytro camera technology: theory, algorithms, performance analysis. In *IS&T/SPIE Electronic Imaging* (2013), International Society for Optics and Photonics, pp. 86671J–86671J.

[9] Ghasemi, and Vetterli. Vlfeat: An open and portable library of computer vision algorithms. In *Proceedings of IS&T SPIE EI 2014* (2014), SPIE.

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

[11] Griffin, G., Holub, A., and Perona, P. Caltech-256 object category dataset. Tech. Rep. 7694, California Institute of Technology, 2007.

[12] J. Xiao, J. Hays, K. E. A. O., and Torralba, A. "sun database: Large-scale scene recognition from abbey to zoo".

[13] Kevin Lai, Liefeng Bo, X. R., and Fox, D. A large-scale hierarchical multi-view rgb-d object dataset. *IEEE International Conference on Robotics and Automation (ICRA)* (May 2011), 1817 – 1824.

[14] Opelt, A., Fussenegger, M., Pinz, A., and Auer, P. Generic object recognition with boosting. *IEEE Trans. PAMI* (2006).

[15] S. A. Nene, S. K. N., and Murase, H. "columbia object image library (coil-100)", February 1996. Technical Report CUCS-006-96.

[16] Vedaldi, A., and Fulkerson, B. Vlfeat: An open and portable library of computer vision algorithms. In *Proceedings of the international conference on Multimedia* (2010), ACM, pp. 1469–1472.