MIT AI LAB MEMO AIM-2005-025, SEPTEMBER, 2005.

LabelMe: a database and web-based tool for image annotation

BRYAN C. RUSSELL, ANTONIO TORRALBA,
Computer Science and Artificial Intelligence Laboratory,
Massachusetts Institute of Technology, Cambridge, MA 02139, USA
brussell@csail.mit.edu, torralba@csail.mit.edu

KEVIN P. MURPHY
Departments of computer science and statistics,
University of British Columbia, Vancouver, BC V6T 1Z4
murphyk@cs.ubc.ca

WILLIAM T. FREEMAN
Computer Science and Artificial Intelligence Laboratory,
Massachusetts Institute of Technology, Cambridge, MA 02139, USA
billf@csail.mit.edu

Abstract

Research in object detection and recognition in cluttered scenes requires large image collections with ground truth labels. The labels should provide information about the object classes present in each image, as well as their shape and locations, and possibly other attributes such as pose. Such data is useful for testing, as well as for supervised learning. This project provides a web-based annotation tool that makes it easy to annotate images, and to instantly share such annotations with the community. This tool, plus an initial set of 10,000 images (3000 of which have been labeled), can be found at
http://www.csail.mit.edu/∼brussell/research/LabelMe/intro.html

1 Introduction

Detecting generic object categories in natural, cluttered images is one of the holy grails of computer vision. In order to make progress towards this goal, it may be essential to have large databases of challenging images, in which "ground truth" labels are made publically available. These labels should provide information about the object classes present in each image, as well as their shape and locations, and possibly other attributes such as pose. This data can be used for testing (comparing algorithms), as well as for training using supervised learning techniques. We discuss the need for creating labeled training databases in more detail in Section 2 below.

Although large labeled databases are standard in other fields, such as speech recognition, natural language processing, and video retrieval, they are comparatively rare in the object detection community. In Section 3, we summarize some of the existing databases, and explain why we feel they are inadequate for the task of learning to detect thousands of object categories in cluttered images. LabelMe has the potential to overcome those problems.

LabelMe is not only a database of labeled images, but also a web-based tool for easily creating and sharing more labeled data. In Section 4, we discuss existing annotation tools, and why they are unsuitable for our needs. In Section 5, we discuss some details of the database itself, and in Section 6, we conclude.

2 The need for labeled image databases

Recently it has become popular to learn to recognize objects from "weakly labeled" images. This takes two main forms: (1) images with captions, which indicate the objects that are present (e.g., Corel data[1]), and (2) images which are known to contain (or known not to contain) the object of interest (e.g., the Caltech dataset[2]). Examples of methods that can learn from such weakly labeled data include [2], which applies segmentation techniques to the Corel data, and [6], which applies interest point detectors to the Caltech data. Both methods then attempt to solve a hard data association problem. More recently, it has become popular to apply "bag of word" techniques to discover object categories from images which are known to contain the object [10, 5, 12, 11, 9].

The relative success of such methods raises the question of whether we need images with more detailed annotation (which is more labor intensive to acquire than just captions). We argue that detailed annotation is necessary, for several reasons. First, labeled data is necessary to quantitatively measure performance of different methods (the issue of how to compare object detection algorithms is beyond the scope of this paper, but see [1] and [8, p99] for some relevant issues). Second, labeled data is necessary because current segmentation and interest point techniques are not capable of discovering the outlines/shapes of many object categories, which are often small and indistinct in natural images. Some kind of "focus of attention" is necessary. This can take many forms: motion cues, stereo cues, manual labeling, or cropping of the image (as in the Caltech dataset). (Remember, people learn to recognize objects from a very rich multimodal sensory stream!)

The final reason we believe large labeled databases will be useful is by analogy with the speech and language communities, where history has shown that performance increases dramatically when more labeled training data is made available. One can argue that

[1] www.corel.com. Note that this dataset is no longer available, although many researchers in the field have purchased a copy in the past.

[2] http://www.vision.caltech.edu/Image_Datasets/Caltech101/Caltech101.html

this is a limitation of current learning techniques. This is one motivation behind the recent interest in Bayesian approaches to learning (e.g., [4]) and to multi-task learning (e.g., [13]). Nevertheless, even if we can learn each class from just a small number of examples, there are still many classes to learn (Biederman [3] famously estimates that people can recognize about 3000 entry-level object categories.) Many arguments point to the requirement of much data.

3 Comparison with other databases

There are a large number of publically available databases of visual objects. We do not have space to review them all there.[3] However, we give a brief summary of the main features which distinguish the LabelMe database from others.

- Designed for object class recognition as opposed to instance recognition. To recognize an individual object instance, one needs multiple images of the same object under different views, occlusion conditions, etc. To recognize an object class, one needs multiple images of *different* instances of the same class, as well as different viewing conditions. Many object class recognition databases, however, only contain different instances in a canonical pose, and without any clutter.

- Designed for objects embedded in a scene. Many databases consist of small cropped images of object instances.[4] These are suitable for training patch-based object detectors (such as sliding window classifiers), but cannot be used for testing, or for training detectors that exploit contextual cues.

- High quality labeling. Many databases just provide captions, which specify that the object is present somewhere in the image. However, as argued aboved, more detailed information, such as bounding boxes, polygons or segmentation masks, is tremendously helpful.

- Many diverse object classes. Many databases only contain a small number of classes, such as faces, pedestrians and cars. (A notable exception is the Caltech 101 database, which we discuss below: see Figure 4 for a comparison.)

- Many diverse images. Many datbases only contain images of a certain type, often professional photographs in which the object of interest is centered, or at least in focus. For many applications, it is useful to vary scene type (nature scenes vs. street scenes vs. office scenes), depth of field (landscape shots vs. close-ups), degree of clutter, etc.

- Non-copyrighted images. In the LabelMe database, most of our images were taken by the authors of this paper, using a variety of hand-held digital cameras[5], so the copyright status is clear.

[3] A more detailed summary of existing databases is available at `http://www.cs.ubc.ca/~murphyk/Vision/objectRecognitionDatabases.html`.

[4] http://cbcl.mit.edu/cbcl/software-datasets/index.html

[5] We also have many video sequences taken with a head-mounted web-cam

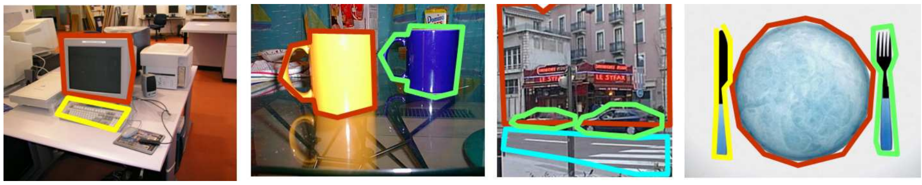


Figure 1: Some examples of images and their corresponding annotations.

- Dynamic, growing database. The LabelMe database is designed to be open-access, and to grow over time.

Below we provide a more detailed comparison of LabelMe with a few of the more popular object databases.

3.1 Comparison with the Caltech databases

The "Caltech Five" database[6] consists of images of cars (rear), motorbikes (side), airplanes (side), faces (front) and leaves (top-down). This database has become quite popular due to its use in an influential series of papers on the constellation model, including [15] and [7]. The problem with this database is that it only contains 5 kinds of objects, and each object more or less completely fills the image, leaving little room for background clutter. This makes it suitable for training/ testing binary image classification (deciding if the image contains the object or is from the generic background class), but not for object detection/ localization. Note that it is now possible to achieve classification performance well above 95% on this dataset, suggesting that it is "too easy", and not a good way to advance the field. The lack of background also makes this dataset not suited for developing algorithms that make use of contextual information for object localization (note that non-contextual location can still be done if one makes a mosaic of images by putting together a 3x3 image array where one contains the target; however, this will have artifacts and will not be useful for contextual analysis).

More recently, Fei-Fei Li *et al*. have collected the "Caltech 101" database [7]. This contains images of 101 object categories (from accordions to yin-yang symbols), collected using Google's image search. There are about 40 to 800 images per category. Like the Caltech-5, this database is suitable for training/testing category level recognition, but not for training/testing object *detection*.

3.2 Comparison with the PASCAL databases

Recently, an EU project called the Visual Object Classes (VOC) challenge (part of the PASCAL network), collected a set of existing databases and put their annotation into a

[6] http://www.vision.caltech.edu/html-files/archive.html

[7] http://www.vision.caltech.edu/Image_Datasets/Caltech101/Caltech101.html

common format [8]. VOC incorporates the MIT CSAIL Database of objects and scenes[9], as well as other databases. LabelMe also incorporates the MIT CSAIL Database of objects and scenes, but has added some additional images/ annotations (in the same general style). In contrast with the PASCAL databases, LabelMe are that LabelMe provides a web-based annotation tool, it uses the XML file format for annotation, and it provides sophisticated Matlab tools for querying and manipulating images and their annotations.

4 Comparison with ESP and Peekaboom

Most existing object databases have been manually labeled using unpublished software. Since our goal is to get other people to label images to ensure that the database increases constantly, we have made our annotation tool (which runs on almost any web browser, and uses a standard Javascript drawing tool[10]) easy to use (see Figure 2 for a screenshot) and freely available. The resulting labels are stored in the XML file format. (See Figure 3 for an example.) This makes the annotations portable and easy to extend.

In order to motivate people to label images, we only make the data available to people who have labeled a minimal number of images on the web site. To make this a bit more fun, we introduce an element of chance: for every object that is labeled, a random number is drawn; if you get a winning number, a new link will appear which will give you access to all of the annotations and images. (The system is currently set up so that, on average, you must label 10 images to get access to the database.) Whenever you want to get an updated version, you must play the game again. In this way, we hope to ensure that the database will continue to grow.

This interactive aspect of the project was inspired by the popularity of the ESP game[11] which has collected over 10 million image captions since 2003. (The images are randomly drawn from the web.) The ESP game requires two users, who view the same image on a web site, to try to "read the mind" of their partner, and to name the object which their partner is currently looking at; however, they are not allowed to use words that are on the taboo list (e.g., previously used words). The person who first reads the other's mind wins that round; thus there is an incentive to label as many images as quickly as possible, with the other person providing quality control. (This has been called "stealing cycles from humans" [14].)

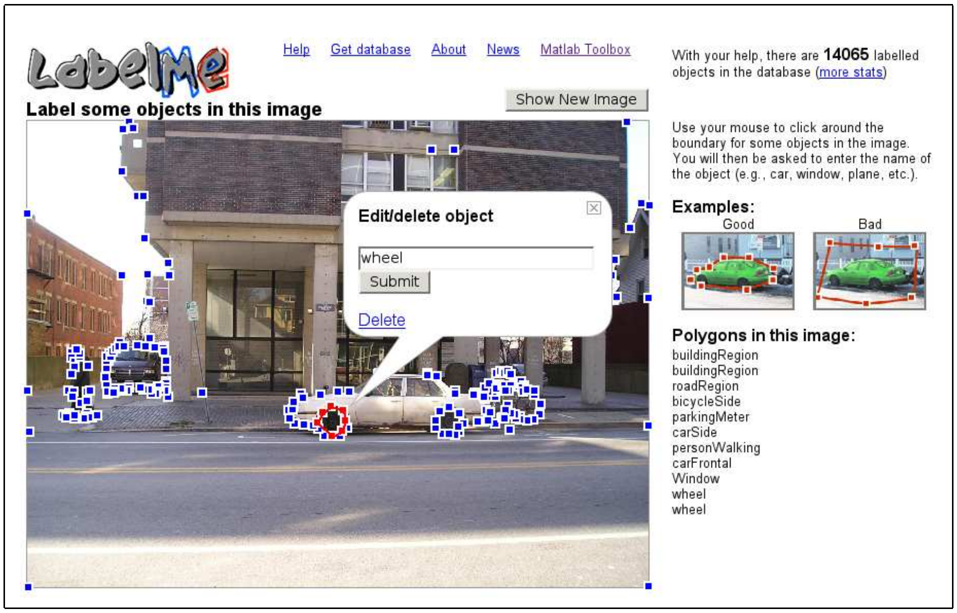
LabelMe is arguably not as fun to use as ESP, and is certainly not as large. However, the quality of the LabelMe data is much higher: the LabelMe images are high resolution images of everyday objects in natural settings, whereas many ESP images are logos or

[8] http://www.pascal-network.org/challenges/VOC/

[9] http://web.mit.edu/torralba/www/database.html

[10] http://www.walterzorn.com/jsgraphics/jsgraphics_e.htm

[11] The ESP game is at www.espgame.org. A search engine, at http://www.captcha.net/esp-search.html, provides access to about 30,000 of these images. Another search engine, at http://hunch.net/∼jl/ESP-ImageSet/search.shtml, provides access to a larger subset (about 68,000 images). In addition, these images and captions can be downloaded from http://hunch.net/∼learning/ESP-ImageSet.tar.gz.



Label some objects in this image

wheel

With your help, there are **14065** labelled objects in the database (more stats)

Use your mouse to click around the boundary for some objects in the image. You will then be asked to enter the name of the object (e.g., car, window, plane, etc.).

Polygons in this image:
buildingRegion
buildingRegion
roadRegion
bicycleSide
parkingMeter
carSide
personWalking
carFrontal
Window
wheel
wheel

Figure 2: Screenshot from the labeling tool in use. The user is shown an image, with possibly one or more existing annotations in the image. The user has the option of annotating a new object, by clicking around the boundary of the object, or editing an existing "unverified" annotation. The user can annotate as many objects in the image as they wish. One finished, the user then clicks the "Show New Image" button to see a new image.

symbolic depictions of isolated objects, and are only available as small images. Also, LabelMe provides polygonal outlines, whereas ESP just provides captions.

More recently, the team that created ESP has created Peekaboom[12]. This is also a two-player game, but now, one player (called Boom) sees an entire image, and must click on little pieces of it to reveal them to the other player (called Peek), who must guess the name of the object. Once guessed, the players switch roles. The goal is to guess as many images as possible within four minutes.

The advantage of Peeakboom over ESP is that it provides location information about the objects of interest, not just captions. However, the data is still of lower quality than LabelMe's, since it does not contain complete object outlines (just little pieces), and the images are often untypical of everyday visual scenes.

[12] www.peekaboom.org

```
<annotation>
 <filename>p1010783.jpg</filename>
 <folder>05june05-static-street-boston</folder>
  <source>
    <sourceImage>
      The MIT-CSAIL database of objects and scenes
    </sourceImage>
    <sourceAnnotation>
      LabelMe Web tool
    </sourceAnnotation>
  </source>
  <object>
   <name>car</name>
   <deleted>0</deleted>
   <verified>1</verified>
    <date>05-Jun-2005 12:12:39</date>
    <polygon>
      <pt> <x>1562</x> <y>1405</y> </pt>
      <pt> <x>1537</x> <y>1426</y> </pt>
      <pt> <x>1557</x> <y>1436</y> </pt>
    </polygon>
  </object>
</annotation>
```

Figure 3: An example XML annotation file for an image containing one object. The polygon defining the object outline has three corners.

5 The LabelMe database

The LabelMe database contains all of the images from the MIT CSAIL Database of objects and scenes[13], in addition to a significant number of newly added images. Altogether, the set consists mostly of office and street scenes.

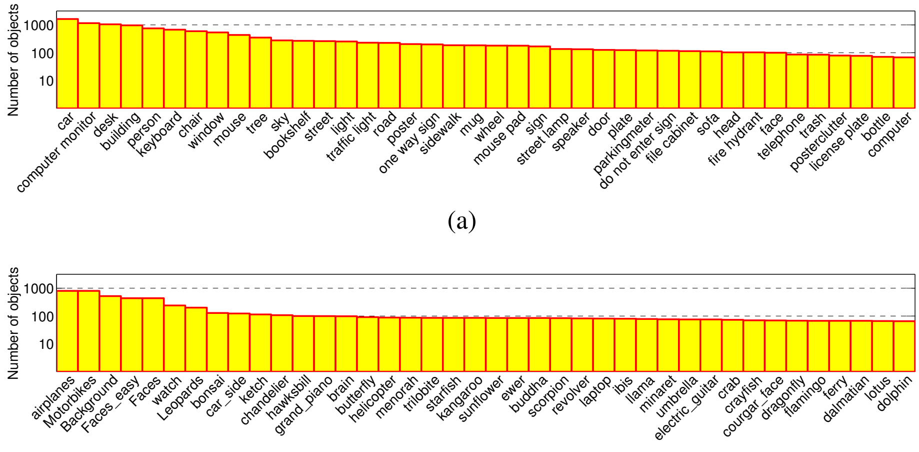
Currently, LabelMe contains 455 distinct object names, and the number of images with at least one labeled object is 3342. Note that these numbers are changing as new polygons get added every week. Figure 4 shows some of the more populated object categories and compares with the Caltech 101 dataset. Notice that our dataset contains on average many more examples per object category.

5.1 Matlab query tools

Since the annotations are in XML format, they are easy to parse using a variety of languages. However, since Matlab is the programming environment of choice for many vision and learning researchers (including the authors), we have created some utilities to convert the XML files into a Matlab data structure. We provide a variety of tools for manipulating this structure, which we summarize below.

- Reading and displaying an annotated image.

[13] http://web.mit.edu/torralba/www/database.html



(a)

(b)

Figure 4: Bar plot showing the top 40 most populated object descriptions for the (a) LabelMe and (b) Caltech 101 (plus background) datasets. The vertical axis is plotted on a logarithmic scale. Notice that the LabelMe dataset contains, on average, more images per object description than the Caltech 101 dataset. The median number of images over the top 40 descriptions is 195 and 85.5 for LabelMe and Caltech 101 respectively. Over 102 descriptions, the median number of images is 46.5 and 59 for LabelMe and Caltech 101 respectively.

- Visualizing the database (showing thumbnails, counting the number of labeled objects, etc.).
- Query the database for images which contain a given object.
- Manipulating images (and their annotations) so that they meet certain criteria, e.g., that all objects of interest have a fixed size (say 64x64 pixels).

For details, please see the web page.

5.2 Quality control

Since people are free to store any annotations they like on the web site, there needs to be some quality control. Currently this is performed manually. A field, called “verified”, is attached to each annotated object, and is set to true if the annotations have been checked for quality by one of the authors. In the future, we hope to semi-automate this by incrementally training and applying an object detector to detect poorly labeled images, and/or to suggest locations of unlabeled objects (c.f., the Seville project [14], for incrementally training a Viola-Jones type detector).

A second issue is deciding what to label. For example, should you label a whole

[14] http://caor.ensmp.fr/∼abramson/seville/

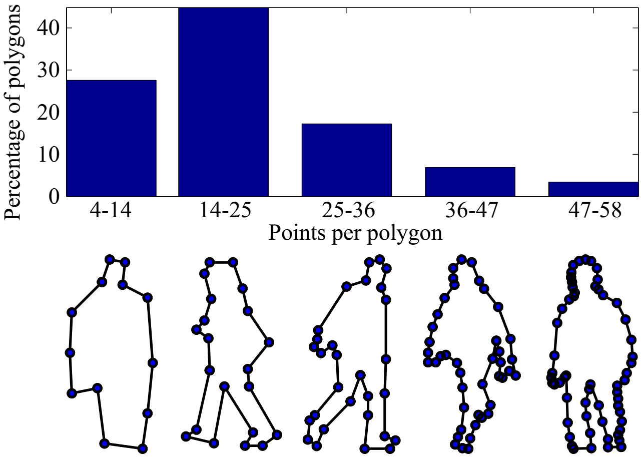


Figure 5: Illustration of the quality of the annotations provided by the users of the annotation tool for the pedestrian object class. The top figure shows the distribution of the number of points in a given polygon and the bottom shows example polygons from each histogram bin. Notice that even the simplest polygons provide a good idea of the object boundary.

pedestrian, or just the head, or just the face? What if it is a crowd of people - should you label all of them? Currently we leave these decisions up to each user. In this way, the annotations will hopefully reflect what various people think are "natural" ways to segment an image (c.f., the Berkeley segmentation database[15], which confronts these same issues in deciding how to segment an image).

A third issue is deciding what description to use. For example, should you call this object a "person", "pedestrian", or "man/woman"? An obvious solution is to provide a drop-down menu of standard object category names. However, we currently prefer to let people use their own descriptions, since these may capture some nuances that will be useful in the future. (For example, people often use the most specific term possible, so they may prefer to say "person" instead of "man/woman" only if they don't know the gender of the person.) In the future we plan to create equivalence classes of labels, using online resources such as Word Net[16].

One important concern with this kind of tool is the quality of the polygons provided by the users. Figure 5 illustrates the statistics of the number of points that define each polygon for the pedestrian object class that were introduced using the web annotation tool. This class is among the most complicated ones. The figure shows that even the simplest polygons provide a good idea of the outline of the object, which is sufficient for most object detection and segmentation algorithms. Notice that there is a significant number of polygons with a high degree of detail.

[15] http://www.cs.berkeley.edu/projects/vision/grouping/segbench

[16] http://wordnet.princeton.edu/

6 Conclusion

We have described the LabelMe dataset and annotation tool. We highlighted the need for a large, rich, and open dataset and showed that LabelMe satisfies many of these demands better than existing datasets. Our dataset will continue to grow through the online annotation tool, with newly added annotations instantly shared. We have provided a Matlab toolbox to parse and query the dataset. We encourage researchers to use the LabelMe dataset and tool, and hope it will advance the field of visual object detection and recognition.

7 Acknowledgements

Financial support was provided by the National Geospatial-Intelligence Agency, NEGI-1582-04-0004, and a grant from BAE Systems. Kevin Murphy was supported in part by a Canadian NSERC Discovery Grant.

References

[1] Shivani Agarwal, Aatif Awan, and Dan Roth. Learning to detect objects in images via a sparse, part-based representation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 26(11):1475–1490, 2004.

[2] Kobus Barnard, Pinar Duygulu, Nando de Freitas, David Forsyth, David Blei, and Michael I. Jordan. Matching words and pictures. *J. of Machine Learning Research*, 3:1107–1135, 2003.

[3] I. Biederman. Recognition by components: a theory of human image interpretation. *Pyschological review*, 94:115–147, 1987.

[4] L. Fei-Fei, R. Fergus, and P. Perona. A bayesian approach to unsupervised one-shot learning of object categories. In *IEEE Intl. Conf. on Computer Vision*, 2003.

[5] L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *CVPR*, 2005.

[6] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *CVPR*, 2003.

[7] R. Fergus, P. Perona, and A. Zisserman. A sparse object category model for efficient learning and exhaustive recognition. In *CVPR*, 2005.

[8] B. Leibe. *Interleaved object categorization and segmentation*. PhD thesis, 2005.

[9] P. Quelhas, F. Monay, J. M. Odobez, D. Gatica-Perez, T. Tuytelaars, and L. Van Gool. Modeling scenes with local descriptors and latent aspects. In *IEEE Intl. Conf. on Computer Vision*, 2005.

[10] J. Sivic, B. Russell, A. Efros, A. Zisserman, and W. Freeman. Discovering objects and their location in images. In *IEEE Intl. Conf. on Computer Vision*, 2005.

[11] E. Sudderth, A. Torralba, W. T. Freeman, and W. Willsky. Describing visual scenes using transformed dirichlet processes. In *Advances in Neural Info. Proc. Systems*, 2005.

[12] E. Sudderth, A. Torralba, W. T. Freeman, and W. Willsky. Learning hierarchical models of scenes, objects, and parts. In *IEEE Intl. Conf. on Computer Vision*, 2005.

[13] A. Torralba, K. Murphy, and W. Freeman. Sharing features: efficient boosting procedures for multiclass object detection. In *CVPR*, 2004.

[14] Luis von Ahn and Laura Dabbish. Labeling images with a computer game. In *Proc. SIGCHI conference on Human factors in computing systems*, 2004.

[15] M. Weber, M. Welling, and P. Perona. Towards automatic discovery of object categories. In *CVPR*, pages 101–109, 2000.