# Thing Detection

## Abstract

## 1. Introduction

### 1.1 Background

Semantic classes in images can be either things (objects with a well-defined shape, e.g. car, person) or stuff (amorphous background regions, e.g. grass, sky) [10]. Often things are also dynamic and able to move, while stuff is more stationary.

The goal of this paper is to present datasets, algorithms and implementations to detect and localize things in an image. The conclusions are made according to how well thing detection is achieved in the context of image-based awareness, Fig. 1.

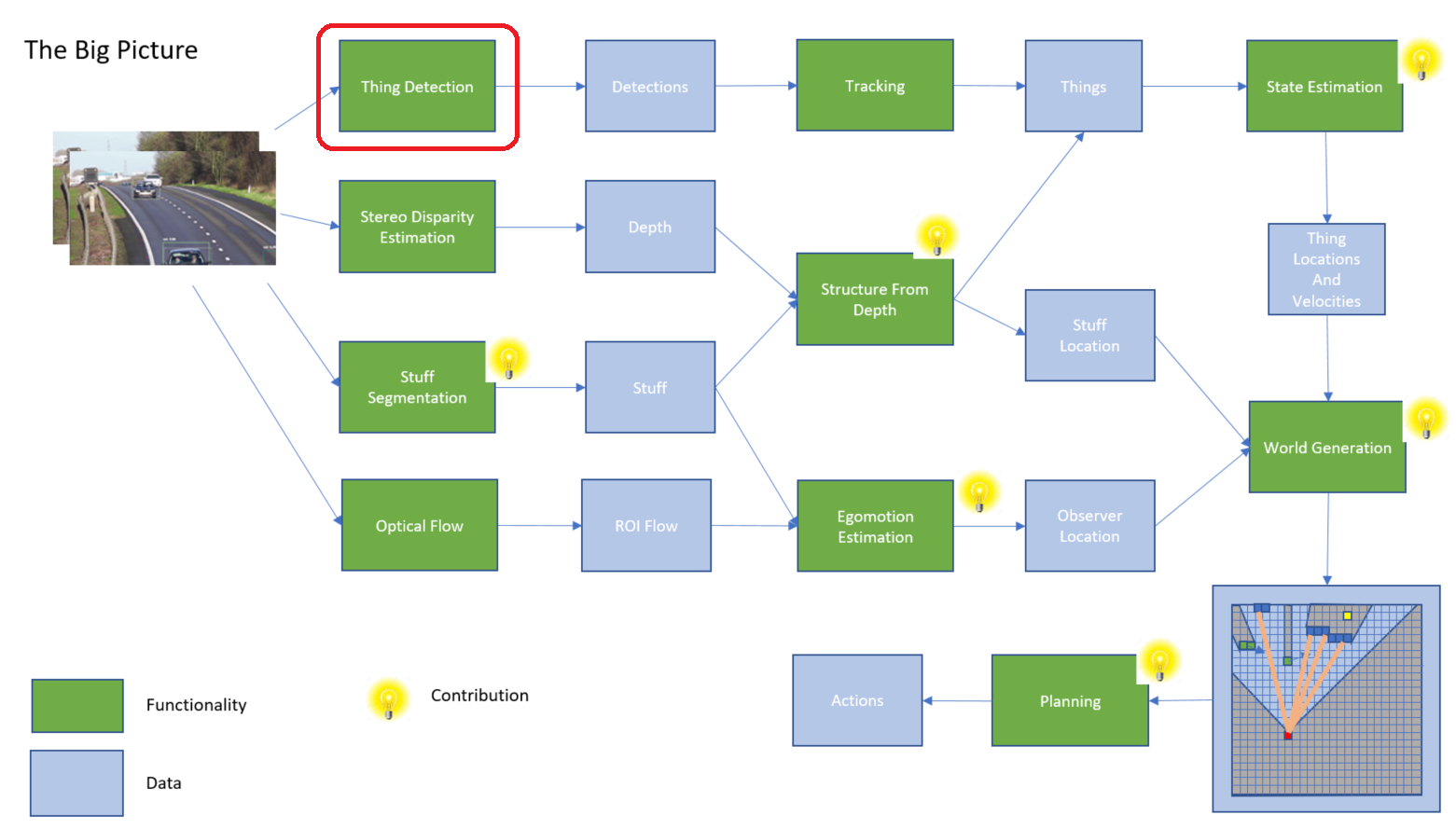


Fig. 1. The role of thing detection in image-based awareness.

### 1.2 Terminology and Tasks

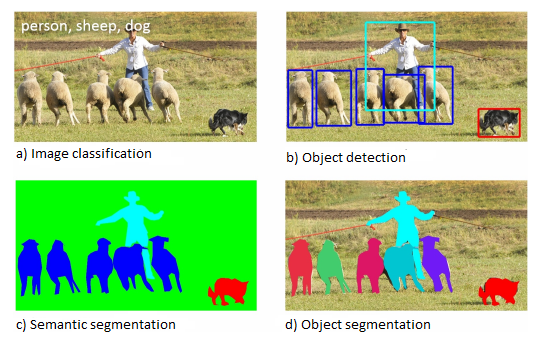
**

Fig. 2. Tasks related to scene understanding.

One of the primary goals of computer vision is the understanding of visual scenes. Scene understanding involves numerous tasks [1]:

**Image classification** requires binary labels indicating whether objects are present in an image.

**Object detection** includes both stating that an object belonging to a specified class is present and localizing it in the image. The location of an object is typically represented by a bounding box.

**Semantic segmentation** requires that each pixel of an image be labelled as belonging to a category. Individual instances of objects do not need to be segmented.

**Object segmentation** combines object detection and semantic segmentation. It requires stating that an object belonging to a specified class is present and localizing it in the image using mask.

**Object recognition** requires binary labels indicating whether specific objects are present in an image, like face recognition for individual persons.

**Instance segmentation** is a synonym for object segmentation.

**Object localization** is object detection for a specific object category.

**Scene recognition** is best defined in [16]. “By ‘scene’ we mean a place in which a human can act within, or a place to which a human being could navigate.”

**Video object segmentation**…

### 1.3 Scope

The scope of this paper is thing detection which is composed of object detection and object segmentation. Object segmentation is more accurate and the preferred task. However, object segmentation also requires more computational power.

## 2. Problem Setup

### 2.1 Input

### 2.2 Output

## 3. Datasets

Selection of dataset determines which categories are available as well as off-the-shelf algorithm implementation. Following common datasets exist (object detection and instance segmentation task related datasets displayed in red):

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Name | Task | Categories | Images | Year | Reference | Notes |
| COIL | Image classification | 20 | 1 440 | 1996 | [6] | Household objects |
| MNIST | Image classification | 10 | 70 000 | 1998 | [5] | Handwritten digits |
| Caltech 101 | Image classification | 101 | 9 146 | 2004 | [7] |  |
| ESP | Image classification | ? (free format labels) | 350 000 | 2004 | [13] | Game driven. Only part of data publicly available (60 000 images) |
| PASCAL VOC (Visual Object Classes) | Object detection,  Instance segmentation | 20 | 11 000 | 2005…2012 | [9] | Several generations |
| MSRC | Image classification | 21 | 591 | 2006 | [11] |  |
| Caltech 256 | Image classification | 256 | 30 607 | 2007 | [8] |  |
| Lotus Hill | Image classification, Object detection, Instance segmentation | ~200 | >50 000 | 2007 | [15] | Available only through purchase |
| The Berkeley Segmentation Data Set | (Semantic) segmentation | ? | 300 | 2007 | [21] | Only partially labelled |
| TinyImage | Image classification | 75 062 | 79 302 107 | 2008 | [12] | 32\*32 colour images |
| LabelMe | Image classification, Object detection, Instance segmentation | ~200 | >30 000 | 2008 | [14] | WordNet classification for part of objects |
| ImageNet | Image classification, Object detection | >20 000  (>1000 for bounding boxes) | 14 197 122  (>1 000 000 for bounding boxes) | 2009… | [2] | WordNet classification structure (synset). No image rights. |
| CIFAR-10 | Image classification | 10 | 60 000 | 2009 | [17] | 32\*32 colour images |
| CIFAR-100 | Image classification | 100 | 60 000 | 2009 | [17] | 32\*32 colour images |
| CamToy (Cambridge-Toyota Labelled Video Database) | Semantic segmentation | 32 | 10 minutes at 1 Hz = 6 000 | 2009 | [20] | Video format |
| SUN | Scene recognition | 899 | 130 519 | 2010 | [16] |  |
| Caltech Pedestrian Dataset | Object localization | 1 | 250 000 | 2012 | [18] |  |
| Indoor Segmentation  Dataset | Semantic segmentation | 26 | 1 449 | 2012 | [19] |  |
| KITTI | Object detection, Instance segmentation, Semantic segmentation | Depends on task | Depends on task | 2012 | [24] | Includes also stereo, depth, 2 and 3d object detection and odometry |
| COCO | Object detection, object segmentation, semantic segmentation | 80 (91) | 330 000 | 2015… | [1], [22] | COCO Stuff was added 2017 |
| Open Images | Object detection | >6000 | >9 000 000 | 2016 | [23] |  |
| CityScapes | Semantic segmentation,  Instance segmentation | 30 | 5 000 (fine), 20 000 (coarse) | 2016 | [25] |  |
| DAVIS: Densely Annotated VIdeo Segmentation | Video instance segmentation | 50 (number of different objects) | 10 459 | 2017 | [26] | Video format |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |

Several other application specific datasets exist, for example to be used in face and pedestrian detection.

## 4. Algorithms

|  |  |  |  |
| --- | --- | --- | --- |
| Name | Task | Reference | Notes |
| YOLO |  |  |  |
| RCNN |  |  |  |
| Fast RCNN |  |  |  |
| Mask RCNN |  |  |  |
| Multibox |  |  |  |
|  |  |  |  |

## 5. Implementations

Following common implementations are currently available:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Algorithm | Dataset | Platform | Model Zoo | Link |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

## 6. Conclusions

## References

[1] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick and Piotr Dollar, “Microsoft COCO: Common Objects in Context”.  [**arXiv:1405.0312**](https://arxiv.org/abs/1405.0312)**.**

[2] 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. [Link](http://www.image-net.org/papers/imagenet_cvpr09.pdf)

[3] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, “The PASCAL visual object classes (VOC) challenge,” IJCV, vol. 88, no. 2, pp. 303–338, Jun. 2010. [Link](http://host.robots.ox.ac.uk/pascal/VOC/pubs/everingham10.pdf)

[4] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, “SUN database: Large-scale scene recognition from abbey to zoo,” in CVPR, 2010. [Link](https://www.cc.gatech.edu/~hays/papers/sun.pdf)

[5] Y. Lecun and C. Cortes, “The MNIST database of handwritten digits,” 1998. [Link](http://yann.lecun.com/exdb/mnist/)

[6] S. A. Nene, S. K. Nayar, and H. Murase, “Columbia object image library (coil-20),” Columbia University, Tech. Rep., 1996. [Link](http://www1.cs.columbia.edu/CAVE/publications/pdfs/Nene_TR96.pdf)

[7] L. Fei-Fei, R. Fergus, and P. Perona, “Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories,” in CVPR Workshop of Generative Model Based Vision (WGMBV), 2004. [Link](https://pdfs.semanticscholar.org/f6ff/fe049408c7b0343a1bdcefe0dc3d0256646d.pdf)

[8] G. Griffin, A. Holub, and P. Perona, “Caltech-256 object category dataset,” California Institute of Technology, Tech. Rep. 7694, 2007. [Link](https://authors.library.caltech.edu/7694/1/CNS-TR-2007-001.pdf)

[9] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, “The PASCAL visual object classes (VOC) challenge,” IJCV, vol. 88, no. 2, pp. 303–338, Jun. 2010. [Link](http://host.robots.ox.ac.uk/pascal/VOC/pubs/everingham10.pdf)

[10] Holger Caesar, Jasper Uijlings, Vittorio Ferrari, “COCO-Stuff: Thing and Stuff Classes in Context”,  [**arXiv:1612.03716**](https://arxiv.org/abs/1612.03716)

[11] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In ECCV06, pages I: 1–15, 2006. [Link](https://webdocs.cs.ualberta.ca/~btap/readings/papers/CRF_TextonBoost_ECCV2006.pdf)

[12] A. Torralba, R. Fergus, and W. Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. PAMI, 30(11):1958–1970, November 2008. [Link](http://people.csail.mit.edu/torralba/publications/80millionImages.pdf)

[13] L. von Ahn and L. Dabbish. Labeling images with a computer game. In CHI04, pages 319–326, 2004. [Link](https://www2.cs.duke.edu/courses/spring07/cps296.3/ESP.pdf)

[14] B. Russell, A. Torralba, K. Murphy, and W. Freeman. Labelme: A database and web-based tool for image annotation. IJCV, 77(1- 3):157–173, May 2008. [Link](http://people.csail.mit.edu/brussell/research/AIM-2005-025-new.pdf)

[15] B. Yao, X. Yang, and S. Zhu. Introduction to a large-scale general purpose ground truth database: Methodology, annotation tool and benchmarks. In EMMCVPR07, pages 169–183, 2007. [Link](https://pdfs.semanticscholar.org/9d18/13749fcb36351fff850f0391968b62fc73b7.pdf)

[16] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, “SUN database: Large-scale scene recognition from abbey to zoo,” in CVPR, 2010. [Link](https://www.cc.gatech.edu/~hays/papers/sun.pdf)

[17] A. Krizhevsky and G. Hinton, “Learning multiple layers of features from tiny images,” Computer Science Department, University of Toronto, Tech. Rep, 2009. [Link](https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf)

[18] P. Dollar, C. Wojek, B. Schiele, and P. Perona, “Pedestrian detection: An evaluation of the state of the art,” PAMI, vol. 34, 2012. [Link](https://pdfs.semanticscholar.org/1423/b4a807f7c0921e33075f016e93b3e8a1dba2.pdf)

[19] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, “Indoor segmentation and support inference from RGBD images,” in ECCV, 2012. [Link](https://www.microsoft.com/en-us/research/wp-content/uploads/2016/11/shkf_eccv2012.pdf)

[20] G. Brostow, J. Fauqueur, and R. Cipolla, “Semantic object classes in video: A high-definition ground truth database,” PRL, vol. 30, no. 2, pp. 88–97, 2009. [Link](http://mi.eng.cam.ac.uk/~gjb47/tmp/prl08.pdf)

[21] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik, “Contour detection and hierarchical image segmentation,” PAMI, vol. 33, no. 5, pp. 898–916, 2011. [Link](https://www2.eecs.berkeley.edu/Research/Projects/CS/vision/grouping/papers/amfm_pami2010.pdf)

[22] [Holger Caesar, Jasper Uijlings, Vittorio Ferrari,” COCO-Stuff: Thing and Stuff Classes in Context” **arXiv:1612.03716**](https://arxiv.org/abs/1612.03716)

[23] I. Krasin, T. Duerig, N. Alldrin, A. Veit, S. Abu-El-Haija, S. Belongie, D. Cai, Z. Feng, V. Ferrari, V. Gomes, A. Gupta, D. Narayanan, C. Sun, G. Chechik, and K. Murphy. Openimages: A public dataset for large-scale multi-label and multi-class image classification. Dataset available from https://github.com/openimages, 2016.

[24] A. Geiger, P. Lenz, and R. Urtasun, “Are we ready for autonomous driving? The Kitti vision benchmark suite,” in Conference on Computer Vision and Pattern Recognition (CVPR), 2012. [Link](http://www.cvlibs.net/publications/Geiger2013IJRR.pdf)

[25] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The Cityscapes Dataset for Semantic Urban Scene Understanding," in Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016. [**arXiv:1604.01685**](https://arxiv.org/abs/1604.01685)

[26] Pont-Tuset, Jordi; Perazzi, Federico; Caelles, Sergi; Arbeláez, Pablo; Sorkine-Hornung, Alex; Luc Van Gool (2017). "The 2017 DAVIS Challenge on Video Object Segmentation".  [arXiv](https://en.wikipedia.org/wiki/ArXiv):[1704.00675](https://arxiv.org/abs/1704.00675)