Visual Object Classes Challenge 2012 (VOC2012)

PASCAL2
Pattern Analysis, Statistical Modelling and Computational Learning

[click on an image to see the annotation]

For news and updates, see the PASCAL Visual Object Classes Homepage

Mark Everingham

It is with great sadness that we report that Mark Everingham died in 2012. Mark was the key member of the VOC project, and it would have been impossible without his selfless contributions. The VOC workshop at ECCV 2012 was dedicated to Mark's memory. A tribute web page has been set up, and an appreciation of Mark's life and work published.

News

- Nov-13: A leaderboard including significance tests will be soon be introduced for new submissions. See Assessing the Significance of Performance Differences on the PASCAL VOC Challenges via Bootstrapping for a description and a demonstration of the method on VOC2012.
- The PASCAL VOC Evaluation Server is open for submissions
- The VOC series of challenges has now finished. We are grateful to the hundreds of participants that have taken part in the challenges over the years.
- 17-Oct-12: Presentations from the workshop are now being placed online.
- 17-Oct-12: Results from the challenge are now available online.
- 07-Oct-12: Provisional programme for the workshop is now online.
- 01-Oct-12: Preliminary results of the challenge are now available to participants.
- 24-Sep-12: The evaluation server is now closed to submissions for the 2012 challenge.
- 03-Sep-12: The PASCAL VOC Evaluation Server is now open for submissions. Note that results on the validation set can be checked, but test set results will be withheld until after the challenge closing date.
- 25-Jun-12: The test data is now available for download from the evaluation server.
- 21-May-12: The development kit is now available for download. Participants should note that data from VOC2011 has been re-used for some tasks: see VOC2012 vs. VOC2011 for details.
- 21-May-12: The challenge workshop will be held on 12th October 2012 in association with ECCV 2012. The workshop format is different to previous years - see the webpage for details.

- 20-Feb-12: The ImageNet Large Scale Visual Recognition Challenge 2012 (ILSVRC2012) will be run in association with the VOC2012 challenge.

Contents

- Introduction
- Data
- VOC2012 vs. VOC2011
- Development kit
- Test data
- Useful software
- Timetable
- Submission of results
- Best practice (Recommendations on using the training and test data)
- Publication policy
- Citation
- Database rights
- Organizers
- Acknowledgements
- Support
- History and background

Introduction

The main goal of this challenge is to recognize objects from a number of visual object classes in realistic scenes (i.e. not pre-segmented objects). It is fundamentally a supervised learning learning problem in that a training set of labelled images is provided. The twenty object classes that have been selected are:

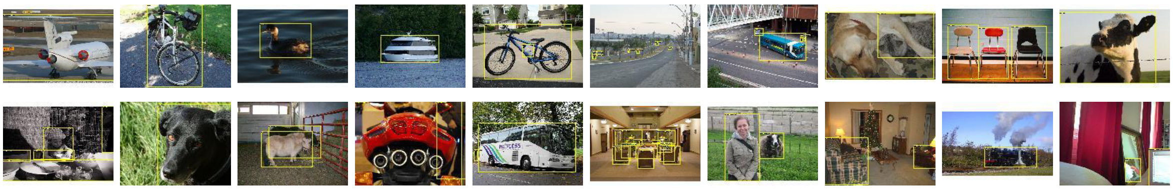
- *Person:* person
- *Animal:* bird, cat, cow, dog, horse, sheep
- *Vehicle:* aeroplane, bicycle, boat, bus, car, motorbike, train
- *Indoor:* bottle, chair, dining table, potted plant, sofa, tv/monitor

There are three main object recognition competitions: classification, detection, and segmentation, a competition on action classification, and a competition on large scale recognition run by ImageNet. In addition there is a "taster" competition on person layout.

Classification/Detection Competitions

1. **Classification**: For each of the twenty classes, predicting presence/absence of an example of that class in the test image.
2. **Detection**: Predicting the bounding box and label of each object from the twenty target classes in the test image.

20 classes



Participants may enter either (or both) of these competitions, and can choose to tackle any (or all) of the twenty object classes. The challenge allows for two approaches to each of the competitions:

1. Participants may use systems built or trained using any methods or data excluding the provided test sets.
2. Systems are to be built or trained using only the provided training/validation data.

The intention in the first case is to establish just what level of success can currently be achieved on these problems and by what method; in the second case the intention is to establish which method is most successful given a specified training set.

Segmentation Competition

- **Segmentation**: Generating pixel-wise segmentations giving the class of the object visible at each pixel, or "background" otherwise.

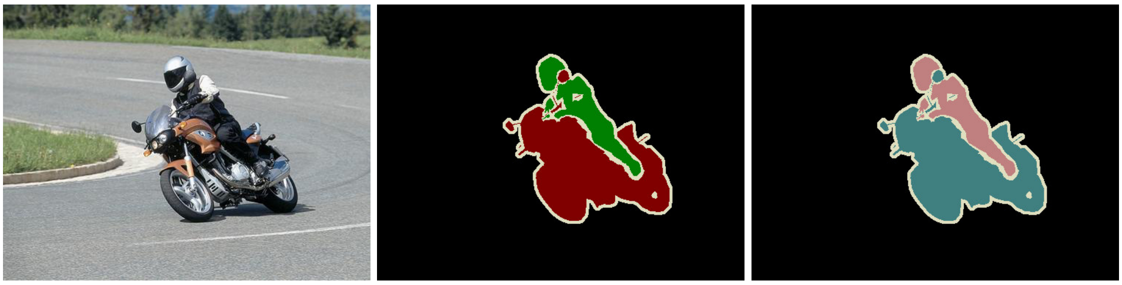
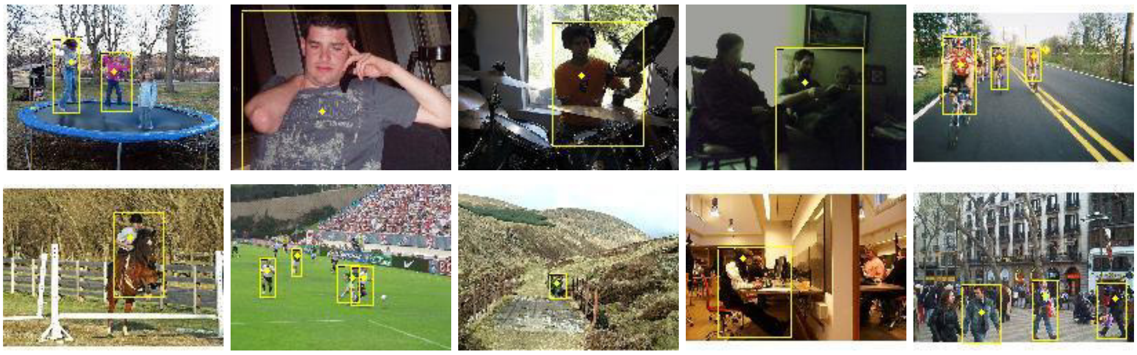


Image Objects Class

Action Classification Competition

- **Action Classification**: Predicting the action(s) being performed by a person in a still image.

10 action classes + "other"



In 2012 there are two variations of this competition, depending on how the person whose actions are to be classified is identified in a test image: (i) by a tight bounding box around the person; (ii) by only a single point located somewhere on the body. The latter competition aims to investigate the performance of methods given only approximate localization of a person, as might be the output from a generic person detector.

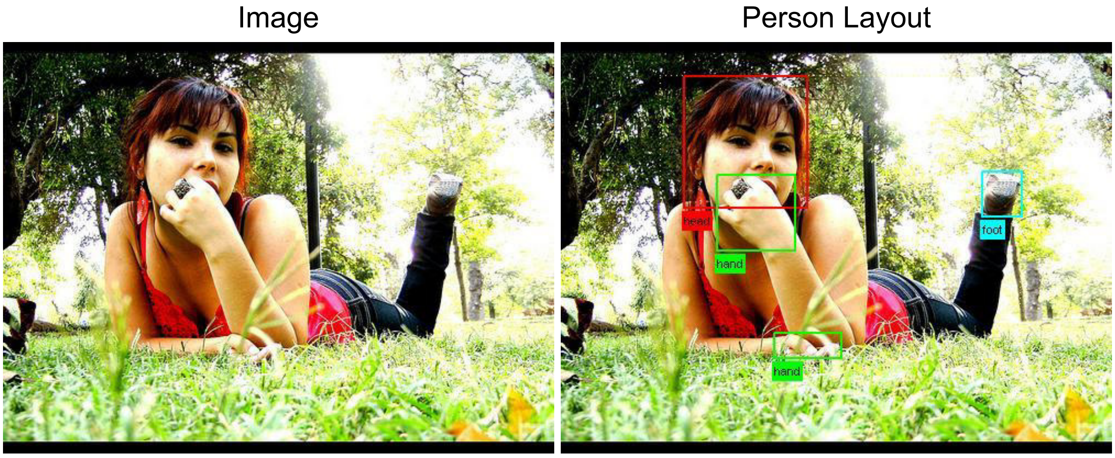
ImageNet Large Scale Visual Recognition Competition

The goal of this competition is to estimate the content of photographs for the purpose of retrieval and automatic annotation using a subset of the large hand-labeled ImageNet dataset (10,000,000 labeled images depicting 10,000+ object categories) as training. Test images will be presented with no initial annotation - no segmentation or labels - and algorithms will have to produce labelings specifying what objects are present in the images. In this initial version of the challenge, the goal is only to identify the main objects present in images, not to specify the location of objects.

Further details can be found at the ImageNet website.

Person Layout Taster Competition

- **Person Layout**: Predicting the bounding box and label of each part of a person (head, hands, feet).



Data

To download the training/validation data, see the development kit.

The training data provided consists of a set of images; each image has an annotation file giving a bounding box and object class label for each object in one of the twenty classes present in the image. Note that multiple objects from multiple classes may be present in the same image. Annotation was performed according to a set of guidelines distributed to all annotators.

A subset of images are also annotated with pixel-wise segmentation of each object present, to support the segmentation competition.

Images for the action classification task are disjoint from those of the classification/detection/segmentation tasks. They have been partially annotated with people, bounding boxes, reference points and their actions. Annotation was performed according to a set of guidelines distributed to all annotators.

Images for the person layout taster, where the test set is disjoint from the main tasks, have been additionally annotated with parts of the people (head/hands/feet).

The data will be made available in two stages; in the first stage, a development kit will be released consisting of training and validation data, plus evaluation software (written in MATLAB). One purpose of the validation set is to demonstrate how the evaluation software works ahead of the competition submission.

In the second stage, the test set will be made available for the actual competition. As in the VOC2008-2011 challenges, no ground truth for the test data will be released.

The data has been split into 50% for training/validation and 50% for testing. The distributions of images and objects by class are approximately equal across the training/validation and test sets. Statistics of the database are online.

Example images

Example images and the corresponding annotation for the classification/detection/segmentation/action tasks, and person layout taster can be viewed online:

- Classification/detection example images
- Segmentation example images
- Action Classification example images
- Person Layout taster example images

VOC2012 vs. VOC2011

For VOC2012 the majority of the annotation effort was put into increasing the size of the segmentation and action classification datasets, and no additional annotation was performed for the classification/detection tasks. The list below summarizes the differences in the data between VOC2012 and VOC2011.

- **Classification/Detection:** The 2012 dataset is the same as that used in 2011. No additional data has been annotated. For this reason, participants are not allowed to run evaluation on the VOC2011 dataset, and this option on the evaluation server has been disabled.
- **Segmentation:** The 2012 dataset contains images from 2008-2011 for which additional segmentations have been prepared. As in previous years the assignment to training/test sets has been maintained. The total number of images with segmentation has been increased from 7,062 to 9,993.
- **Action Classification:** The 2012 dataset comprises the 2011 dataset plus additional annotated images. The assignment to training/test sets has been maintained. In addition to the box annotation, people are now also annotated with a reference point on the body to support the "boxless" action classification task (see the development kit).
- **Person Layout Taster:** The 2012 dataset is the same as that used in 2011. No additional data has been annotated. For this reason, participants are not allowed to run evaluation on the VOC2011 dataset, and this option on the evaluation server has been disabled.

Development Kit

The development kit consists of the training/validation data, MATLAB code for reading the annotation data, support files, and example implementations for each competition.

The development kit is now available:

- Download the training/validation data (2GB tar file)
- Download the development kit code and documentation (500KB tar file)
- Download the PDF documentation (500KB PDF)
- Browse the HTML documentation
- View the guidelines used for annotating the database (VOC2011)
- View the action guidelines used for annotating the action task images

Test Data

The test data will be made available according to the challenge timetable. Note that the only annotation in the data is for the action task and layout taster. As in 2008-2011, there are no current plans to release full annotation - evaluation of results will be provided by the organizers.

The test data can be downloaded from the evaluation server. You can also use the evaluation server to evaluate your method on the test data.

Useful Software

Below is a list of software you may find useful, contributed by participants to previous challenges.

- Encoding Methods Evaluation Toolkit
  Ken Chatfield, Victor Lempitsky, Andrea Vedaldi, Andrew Zisserman

- CPMC: Constrained Parametric Min-Cuts for Automatic Object Segmentation
  Joao Carreira and Cristian Sminchisescu.

- Automatic Labelling Environment (Semantic Segmentation)
  Lubor Ladicky, Philip H.S. Torr.

- Discriminatively Trained Deformable Part Models
  Pedro Felzenszwalb, Ross Girshick, David McAllester, Deva Ramanan.

- Color Descriptors
  Koen van de Sande, Theo Gevers, Cees Snoek.

Timetable

- May 2012: Development kit (training and validation data plus evaluation software) made available.
- 25th June 2012: Test set made available.
- 23rd September 2012 (Sunday, 2300 hours GMT): Deadline for submission of results (there will be no extension).
- 12th October 2012: Challenge workshop in association with ECCV 2012.

Submission of Results

Participants are expected to submit a single set of results per method employed. Participants who have investigated several algorithms may submit one result per method. Changes in algorithm parameters do *not* constitute a different method - all parameter tuning must be conducted using the training and validation data alone.

Results must be submitted using the automated evaluation server:

- PASCAL VOC Evaluation Server

It is essential that your results files are in the correct format. Details of the required file formats for submitted results can be found in the development kit documentation. The results files should be collected in a single archive file (tar/tgz/tar.gz).

Participants submitting results for several different methods (noting the definition of different methods above) should produce a **separate** archive for each method.

In addition to the results files, participants will need to additionally specify:

- contact details and affiliation
- list of contributors
- **description of the method (minimum 500 characters) - see below**

Since 2011 we require all submissions to be accompanied by an abstract describing the method, of minimum length **500 characters**. The abstract will be used in part to select invited speakers at the challenge workshop. If you are unable to submit a description due e.g. to commercial interests or other issues of confidentiality you must contact the organisers to discuss this. Below are two example descriptions, for classification and detection methods previously presented at the challenge workshop. *Note these are our own summaries, not provided by the original authors*.

- **Example Abstract: Object classification**
  Based on the VOC2006 QMUL description of LSPCH by Jianguo Zhang, Cordelia Schmid, Svetlana Lazebnik, Jean Ponce in sec 2.16 of The PASCAL Visual Object Classes Challenge 2006 (VOC2006) Results.

  We make use of a bag-of-visual-words method (cf Csurka et al 2004). Regions of interest are detected with a Laplacian detector (Lindeberg, 1998), and normalized for scale. A SIFT descriptor (Lowe 2004) is then computed for each detection. 50,000 randomly selected descriptors from the training set are then vector quantized (using k-means) into k=3000 "visual words" (300 for each of the 10 classes). Each image is then represented by the histogram of how often each visual word is used. We also make use a spatial pyramid scheme (Lazebnik et al, CVPR 2006). We first train SVM classifiers using the chi^2 kernel based on the histograms of each level in the pyramid. The outputs of these SVM classifiers are then concatenated into a feature vector for each image and used to learn another SVM classifier based on a Gaussian RBF kernel.

- **Example Abstract: Object detection**
  Based on "Object Detection with Discriminatively Trained Part Based Models"; Pedro F. Felzenszwalb, Ross B. Girshick, David McAllester and Deva Ramanan; IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 32, No. 9, September 2010.

  We introduce a discriminatively-trained parts-based model for object detection. The model

> consists of a coarse "root" template of HOG features (Dalal and Triggs, 2006), plus a number of higher-resolution part-based HOG templates which can translate in a neighborhood relative to their default position. The responses of the root and part templates are combined by a latent-SVM model, where the latent variables are the offsets of the parts. We introduce a novel training algorithm for the latent SVM. We also make use of an iterative training procedure exploiting "hard negative" examples, which are negative examples incorrectly classified in an earlier iteration. Finally the model is scanned across the test image in a "sliding-window" fashion at a variety of scales to produce candidate detections, followed by greedy non-maximum suppression. The model is applied to all 20 PASCAL VOC object detection challenges.

If you would like to submit a more detailed description of your method, for example a relevant publication, this can be included in the results archive.

Best Practice

The VOC challenge encourages two types of participation: (i) methods which are trained using only the provided "trainval" (training + validation) data; (ii) methods built or trained using any data except the provided test data, for example commercial systems. In both cases the *test* data must be used strictly for reporting of results alone - it must not be used in any way to train or tune systems, for example by runing multiple parameter choices and reporting the best results obtained.

If using the training data we provide as part of the challenge development kit, all development, e.g. feature selection and parameter tuning, must use the "trainval" (training + validation) set alone. One way is to divide the set into training and validation sets (as suggested in the development kit). Other schemes e.g. *n*-fold cross-validation are equally valid. The tuned algorithms should then be run only *once* on the test data.

In VOC2007 we made all annotations available (i.e. for training, validation and test data) but since then we have not made the test annotations available. Instead, results on the test data are submitted to an evaluation server.

Since algorithms should only be run *once* on the test data we strongly discourage multiple submissions to the server (and indeed the number of submissions for the same algorithm is strictly controlled), as the evaluation server should not be used for parameter tuning.

We encourage you to publish test results always on the latest release of the challenge, using the output of the evaluation server. If you wish to compare methods or design choices e.g. subsets of features, then there are two options: (i) use the entire VOC2007 data, where all annotations are available; (ii) report cross-validation results using the latest "trainval" set alone.

Policy on email address requirements when registering for the evaluation server

In line with the Best Practice procedures (above) we restrict the number of times that the test data can be processed by the evaluation server. To prevent any abuses of this restriction an institutional email address is required when registering for the evaluation server. This aims to prevent one user registering multiple times under different emails. Institutional emails include academic ones, such as name@university.ac.uk, and corporate ones, but not personal ones, such as name@gmail.com or name@123.com.

Publication Policy

The main mechanism for dissemination of the results will be the challenge webpage.

The detailed output of each submitted method will be published online e.g. per-image confidence for the classification task, and bounding boxes for the detection task. The intention is to assist others in the community in carrying out detailed analysis and comparison with their own methods. The published results will not be anonymous - by submitting results, participants are agreeing to have their results shared online.

Citation

If you make use of the VOC2012 data, please cite the following reference (to be prepared after the challenge workshop) in any publications:

```
@misc{pascal-voc-2012,
	author = "Everingham, M. and Van~Gool, L. and Williams, C. K. I. and Winn, J. and Zisserman, A.",
	title = "The {PASCAL} {V}isual {O}bject {C}lasses {C}hallenge 2012 {(VOC2012)} {R}esults",
	howpublished = "http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html"}
```

Database Rights

The VOC2012 data includes images obtained from the "flickr" website. Use of these images must respect the corresponding terms of use:

- "flickr" terms of use

For the purposes of the challenge, the identity of the images in the database, e.g. source and name of owner, has been obscured. Details of the contributor of each image can be found in the annotation to be included in the final release of the data, after completion of the challenge. Any queries about the use or ownership of the data should be addressed to the organizers.

Organizers

- Mark Everingham (University of Leeds)
- Luc van Gool (ETHZ, Zurich)
- Chris Williams (University of Edinburgh)
- John Winn (Microsoft Research Cambridge), john@johnwinn.org
- Andrew Zisserman (University of Oxford)

Acknowledgements

We gratefully acknowledge the following, who spent many long hours providing annotation for the VOC2012 database:

Yusuf Aytar, Lucia Ballerini, Hakan Bilen, Ken Chatfield, Mircea Cimpoi, Ali Eslami, Basura Fernando, Christoph Godau, Bertan Gunyel, Phoenix/Xuan Huang, Jyri Kivinen, Markus Mathias, Kristof Overdulve, Konstantinos Rematas, Johan Van Rompay, Gilad Sharir, Mathias Vercruysse, Vibhav Vineet, Ziming Zhang, Shuai Kyle Zheng.

We also thank Yusuf Aytar for continued development and administration of the evaluation server, and Ali Eslami for analysis of the results.

Support

The preparation and running of this challenge is supported by the EU-funded PASCAL2 Network of Excellence on Pattern Analysis, Statistical Modelling and Computational Learning.

History and Background

The main challenges have run each year since 2005. For more background on VOC, the following journal paper discusses some of the choices we made and our experience in running the challenge, and gives a more in depth discussion of the 2007 methods and results:

The PASCAL Visual Object Classes (VOC) Challenge
Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J. and Zisserman, A.
International Journal of Computer Vision, 88(2), 303-338, 2010
Bibtex source | Abstract | PDF

The table below gives a brief summary of the main stages of the VOC development.

Year	Statistics	New developments	Notes
2005	Only 4 classes: bicycles, cars, motorbikes, people. Train/validation/test: 1578 images containing 2209 annotated objects.	Two competitions: classification and detection	Images were largely taken from exising public datasets, and were not as challenging as the flickr images subsequently used. This dataset is obsolete.
2006	10 classes: bicycle, bus, car, cat, cow, dog, horse, motorbike, person, sheep. Train/validation/test: 2618 images containing 4754 annotated objects.	Images from flickr and from Microsoft Research Cambridge (MSRC) dataset	The MSRC images were easier than flickr as the photos often concentrated on the object of interest. This dataset is obsolete.
2007	20 classes: • *Person:* person • *Animal:* bird, cat, cow, dog, horse, sheep • *Vehicle:* aeroplane, bicycle, boat, bus, car, motorbike, train • *Indoor:* bottle, chair, dining table, potted plant, sofa, tv/monitor Train/validation/test: 9,963 images containing 24,640 annotated objects.	• Number of classes increased from 10 to 20 • Segmentation taster introduced • Person layout taster introduced • Truncation flag added to annotations • Evaluation measure for the classification challenge changed to Average Precision. Previously it had been ROC-AUC.	This year established the 20 classes, and these have been fixed since then. This was the final year that annotation was released for the testing data.
2008	20 classes. The data is split (as usual) around 50% train/val and 50% test. The train/val data has 4,340 images containing 10,363 annotated objects.	• Occlusion flag added to annotations. • Test data annotation no longer made public. • The segmentation and person layout data sets include images from the corresponding VOC2007 sets.	
2009	20 classes. The train/val data has 7,054 images containing 17,218 ROI annotated objects and 3,211 segmentations.	• From now on the data for all tasks consists of the previous years' images augmented with new images. In earlier years an entirely new data set was released each year for the	• No difficult flags were provided for the additional images (an omission). • Test data annotation not

		classification/detection tasks. • Augmenting allows the number of images to grow each year, and means that test results can be compared on the previous years' images. • Segmentation becomes a standard challenge (promoted from a taster)	made public.
2010	20 classes. The train/val data has 10,103 images containing 23,374 ROI annotated objects and 4,203 segmentations.	• Action Classification taster introduced. • Associated challenge on large scale classification introduced based on ImageNet. • Amazon Mechanical Turk used for early stages of the annotation.	• Method of computing AP changed. Now uses all data points rather than TREC style sampling. • Test data annotation not made public.
2011	20 classes. The train/val data has 11,530 images containing 27,450 ROI annotated objects and 5,034 segmentations.	• Action Classification taster extended to 10 classes + "other".	• Layout annotation is now not "complete": only people are annotated and some people may be unannotated.
2012	20 classes. The train/val data has 11,530 images containing 27,450 ROI annotated objects and 6,929 segmentations.	• Size of segmentation dataset substantially increased. • People in action classification dataset are additionally annotated with a reference point on the body.	• Datasets for classification, detection and person layout are the same as VOC2011.