Everingham, M., Gool, L., Williams, C. K. I., Winn, J., & Zisserman, A. (2010). The PASCAL Visual Object Classes (VOC) Challenge. International Journal of Computer Vision, 88(2), 303-338, doi: 10.1007/s11263-009-0275-4

The PASCAL Visual Object Classes (VOC) Challenge

Mark Everingham · **Luc Van Gool** · **Christopher K. I. Williams** · **John Winn** · **Andrew Zisserman**

Received: date / Accepted: date

Abstract The PASCAL Visual Object Classes (VOC) challenge is a benchmark in visual object category recognition and detection, providing the vision and machine learning communities with a standard dataset of images and annotation, and standard evaluation procedures. Organised annually from 2005 to present, the challenge and its associated dataset has become accepted as *the* benchmark for object detection.

This paper describes the dataset and evaluation procedure. We review the state-of-the-art in evaluated methods for both classification and detection, analyse whether the methods are statistically different, what they are learning from the images (e.g. the object or its context), and what the methods find easy or confuse. The paper concludes with lessons learnt in the three year history of the challenge, and proposes directions for future improvement and extension.

Keywords Database · Benchmark · Object recognition · Object detection

M. Everingham
University of Leeds, UK,
E-mail: m.everingham@leeds.ac.uk

L. Van Gool
KU Leuven, Belgium

C.K.I. Williams
University of Edinburgh, UK

J. Winn
Microsoft Research, Cambridge, UK

A. Zisserman
University of Oxford, UK

1 Introduction

The PASCAL[1] Visual Object Classes (VOC) Challenge consists of two components: (i) a publicly available *dataset* of images and annotation, together with standardised evaluation software; and (ii) an annual *competition* and workshop. The VOC2007 dataset consists of annotated consumer photographs collected from the flickr[2] photo-sharing web-site. A new dataset with ground truth annotation has been released each year since 2006. There are two principal challenges: *classification* – "does the image contain any instances of a particular object class?" (where the object classes include cars, people, dogs, etc), and *detection* – "where are the instances of a particular object class in the image (if any)?". In addition, there are two subsidiary challenges ("tasters") on pixel-level segmentation – assign each pixel a class label, and "person layout" – localise the head, hands and feet of people in the image. The challenges are issued with deadlines each year, and a workshop held to compare and discuss that year's results and methods. The datasets and associated annotation and software are subsequently released and available for use at any time.

The objectives of the VOC challenge are twofold: first to provide challenging images and high quality annotation, together with a standard evaluation methodology – a "plug and play" training and testing harness so that performance of algorithms can be compared (the dataset component); and second to measure the state of the art each year (the competition component).

The purpose of this paper is to describe the challenge: what it is, and the reasons for the way it is. We also describe the methods, results and evaluation of the challenge, and so

[1] PASCAL stands for pattern analysis, statistical modelling and computational learning. It is an EU Network of Excellence funded under the IST Programme of the European Union.

[2] http://www.flickr.com/

in that respect are describing the state-of-the-art in object recognition (at least as measured for these challenges and by those who entered). We focus mainly on the 2007 challenge, as this is the most recent, but also discuss significant changes since earlier challenges and why these were made.

1.1 Relation to other Datasets

Challenge datasets are important in many areas of research in order to set goals for methods, and to allow comparison of their performance. Similar datasets and evaluation methodologies are sprouting in other areas of computer vision and machine learning, e.g. the Middlebury datasets for stereo, MRF optimisation, and optical flow comparison (Scharstein and Szeliski 2002).

In addition to organised challenges, there are several datasets contributed by the vision community which are related to that collected for the VOC challenges.

The "Caltech 101" dataset (Fei-Fei et al 2006) contains images of 101 categories of object, and is relatively widely used within the community for evaluating object recognition. Each image contains only a single object. A principal aim of the Caltech datasets is to evaluate multi-category object recognition, as a function of the (relatively small) number of training images. This is complementary to the aims of the VOC challenge, which measures performance on a smaller number of classes and without such constraints on the amount of training data available.

A common criticism of this dataset, addressed by the VOC challenge, is that the images are largely without clutter, variation in pose is limited, and the images have been manually aligned to reduce the variability in appearance. These factors make the dataset less applicable to "real world" evaluation than the images provided for the VOC challenge.

The "Caltech 256" dataset (Griffin et al 2007) corrected some of the deficiencies of Caltech 101 – there is more variability in size and localisation, and obvious artifacts have been removed. The number of classes is increased (from 101 to 256) and the aim is still to investigate multi-category object recognition with a limited number of training images. For the most part there is only a single object per image – as is required to support the 1-of-*m* evaluation adopted ("which one of *m* classes does this image contain?").

The "LabelMe" dataset (Russell et al 2008) at MIT is most similar to the VOC challenge dataset in that it contains more-or-less general photographs containing multiple objects. LabelMe has been ground-breaking in providing a web-based annotation interface, encouraging casual and professional users alike to contribute and share annotation. Many object categories are labelled, with annotation consisting of a bounding polygon and category, with some objects additionally being labelled with pose and object parts. For the most part the dataset is *incompletely* labelled – volunteers are free to choose which objects to annotate, and which to omit. This means that, while a very valuable resource for training images, the dataset is unsuitable for testing in the manner of the VOC challenge since precision and recall cannot accurately be estimated. Recently the LabelMe organisers have proposed subsets of the database to use for training and testing, which are completely annotated with a set of seven object (person, car) and "stuff" (building, sky, etc.) classes. However, no evaluation protocol is specified.

The "TREC Video Retrieval Evaluation" (TRECVID[3], Smeaton et al (2006)) is also similar to the VOC challenge in that there is a new dataset and competition each year, though the dataset is only available to participants and is not publicly distributed. TRECVID includes several tasks, but the one most related to VOC is termed "high-level feature extraction", and involves returning a ranked list of video shots for specified "features". For the 2008 competition these features include scene categories (such as classroom, cityscape or harbour), object categories (such as dog, aeroplane flying or telephone) and actions/events (such as a demonstration/protest). Annotation is not provided by the organisers, but some is usually distributed amongst the participants. The submissions are scored by their Average Precision (AP). The evaluation of the ranked lists is carried out by the organisers using a mixture of ground truth labelling and "inferred ground truth" (Yilmaz and Aslam 2006) obtained from high ranked results returned by the participants' methods.

The Lotus Hill dataset (Yao et al 2007) is a large, recently produced dataset, a small part of which is made freely available to researchers. It contains 8 data subsets with a range of annotation. Particularly we highlight (a) annotations providing a hierarchical decomposition of individual objects e.g. vehicles (9 classes, 209 images[4]), other man-made objects (75 classes, 750 images) and animals (40 classes, 400 images); and (b) segmentation labelling of scenes to a pixel level (444 images). As this dataset has only recently been released there has not yet been a lot of work reported on it. The datasets look to have a useful level of annotation (especially with regard to hierarchical decompositions which have not been attempted elsewhere), but are somewhat limited by the number of images that are freely available.

[3] `http://www-nlpir.nist.gov/projects/trecvid/`

[4] The number of images quoted is the number that are freely available.

1.2 Paper layout

This paper is organised as follows: we start with a summary of the four challenges in Sect. 2, then describe in more detail in Sect. 3 the datasets – their method of collection; the classes included and the motivation for including them; and their annotation and statistics. Sect. 4 describes the evaluation procedure and why this procedure was chosen. Sect. 5 overviews the main methods used in the 2007 challenge for classification and detection, and Sect. 6 reports and discusses the results. This discussion includes an analysis of the statistical significance of the performances of the different methods, and also of which object classes and images the methods find easy or difficult. We conclude with a discussion of the merits, and otherwise, of the VOC challenge and possible options for the future.

2 Challenge Tasks

This section gives an overview of the two principal challenge tasks on *classification* and *detection*, and on the two subsidiary tasks ("tasters") on pixel-level segmentation, and "person layout".

2.1 Classification

For each of twenty object classes, predict the presence/absence of at least one object of that class in a test image. Participants are required to provide a real-valued confidence of the object's presence for each test image so that a precision/recall curve can be drawn. Participants may choose to tackle all, or any subset of object classes, for example "cars only" or "motorbikes and cars".

Two competitions are defined according to the choice of training data: (1) taken from the VOC training/validation data provided, or (2) from any source excluding the VOC test data. In the first competition, any annotation provided in the VOC training/validation data may be used for training, for example bounding boxes or particular views e.g. "frontal" or "left". Participants are *not* permitted to perform additional manual annotation of either training or test data. In the second competition, any source of training data may be used *except* the provided test images. The second competition is aimed at researchers who have pre-built systems trained on other data, and is a measure of the state-of-the-art.

2.2 Detection

For each of the twenty classes, predict the bounding boxes of each object of that class in a test image (if any), with associated real-valued confidence. Participants may choose to tackle all, or any subset of object classes. Two competitions are defined in a similar manner to the classification challenge.

2.3 Segmentation Taster

For each test image, predict the object class of each pixel, or "background" if the object does not belong to one of the twenty specified classes. Unlike the classification and detection challenges there is only one competition, where training data is restricted to that provided by the challenge.

2.4 Person Layout Taster

For each "person" object in a test image (if any), detect the person, predicting the bounding box of the person, the presence/absence of parts (head/hands/feet), and the bounding boxes of those parts. Each person detection should be output with an associated real-valued confidence. Two competitions are defined in a similar manner to the classification challenge.

3 Datasets

The goal of the VOC challenge is to investigate the performance of recognition methods on a wide spectrum of natural images. To this end, it is required that the VOC datasets contain significant variability in terms of object size, orientation, pose, illumination, position and occlusion. It is also important that the datasets do not exhibit systematic bias, for example, favouring images with centred objects or good illumination. Similarly, to ensure accurate training and evaluation, it is necessary for the image annotations to be consistent, accurate and exhaustive for the specified classes. This section describes the processes used for collecting and annotating the VOC2007 datasets, which were designed to achieve these aims.

3.1 Image Collection Procedure

For the 2007 challenge, all images were collected from the flickr photo-sharing web-site. The use of personal photos which were not taken by, or selected by, vision/machine learning researchers results in a very "unbiased" dataset, in the sense that the photos are not taken with a particular purpose in mind i.e. object recognition research. Qualitatively the images contain a very wide range of viewing conditions (pose, lighting, etc.) and images where there is little bias toward images being "of" a particular object, e.g. there are images of motorcycles in a street scene, rather than solely im-

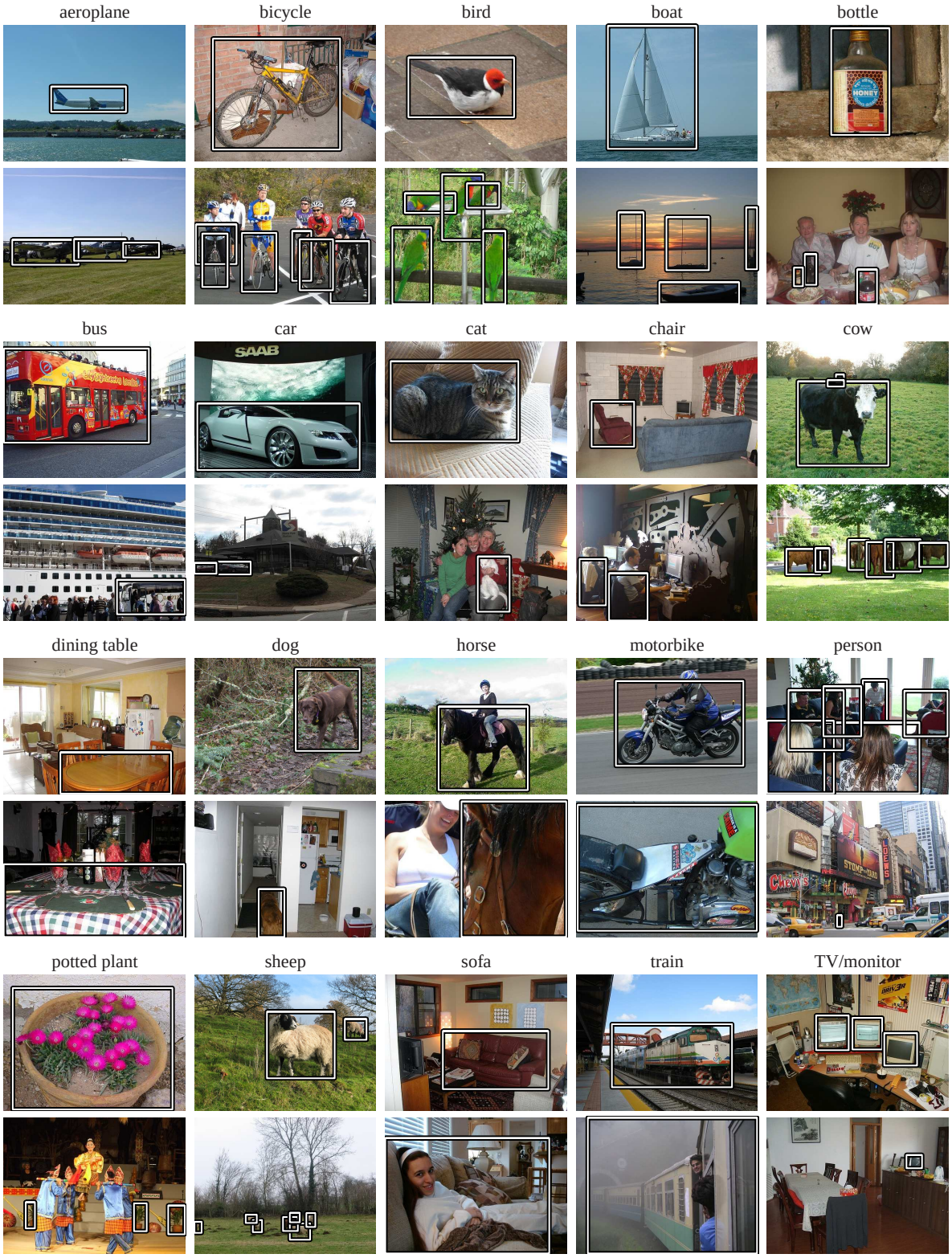


Fig. 1 Example images from the VOC2007 dataset. For each of the 20 classes annotated, two examples are shown. Bounding boxes indicate all instances of the corresponding class in the image which are marked as "non-difficult" (see Sect. 3.3) – bounding boxes for the other classes are available in the annotation but not shown. Note the wide range of pose, scale, clutter, occlusion and imaging conditions.

ages where a motorcycle is the focus of the picture. The annotation guidelines (Winn and Everingham 2007) provided guidance to annotators on which images to annotate – essentially everything which could be annotated with confidence. The use of a single source of "consumer" images addressed problems encountered in previous challenges, such as in VOC2006 where images from the Microsoft Research Cambridge database (Shotton et al 2006) were included. The MSR Cambridge images were taken with the purpose of capturing particular object classes, so that the object instances tend to be large, well-illuminated and central. The use of an *automated* collection method also prevented any selection bias being introduced by a researcher manually performing image selection. The "person" category provides a vivid example of how the adopted collection methodology leads to high variability; in previous datasets "person" was essentially synonymous with "pedestrian", whereas in the VOC dataset we have images of people engaged in a wide range of activities such as walking, riding horses, sitting on buses, etc. (see Fig. 1).

In total, 500,000 images were retrieved from flickr. For each of the 20 object classes to be annotated (see Fig. 1), images were retrieved by querying flickr with a number of related keywords (Table 1). No other query criteria, e.g. date of capture, photographer's name, etc. were specified – we return to this point below.

For a given query, flickr is asked for 100,000 matching images (flickr organises search results as "pages" i.e. 100 pages of 1,000 matches). An image is chosen at random from the returned set and downloaded along with the corresponding metadata. A new query is then selected at random, and the process is repeated until sufficient images have been downloaded. Images were downloaded for each class in parallel using a python interface to the flickr API, with no restriction on the number of images per class or query. Thanks to flickr's fast servers, downloading the entire image set took just a few hours on a single machine.

Table 1 lists the queries used for each of the classes, produced by "free association" from the target classes. It might appear that the use of keyword queries would bias the images to pictures "of" an object, however the wide range of keywords used reduces this likelihood; for example the query "living room" can be expected to return scenes containing chairs, sofas, tables, etc. *in context*, or the query "town centre" to return scenes containing cars, motorcycles, pedestrians, etc. It is worth noting, however, that without using any keyword queries the images retrieved randomly from flickr were, subjectively, found to be overwhelmingly "party" scenes containing predominantly people. We return to the problem of obtaining sufficient examples of "minority" object classes in Sect. 7.1.

All exact duplicate and "near duplicate" images were removed from the downloaded image set, using the method of (Chum et al 2007). Near duplicate images are those that are perceptually similar, but differ in their levels of compression, or by small photometric distortions or occlusions for example.

After de-duplication, random images from the set of 500,000 were presented to the annotators for annotation. During the annotation event, 44,269 images were considered for annotation, being either annotated or discarded as unsuitable for annotation e.g. containing no instances of the 20 object classes, according to the annotation guidelines (Winn

Table 1 Queries used to retrieve images from flickr. Words in bold show the "targeted" class. Note that the query terms are quite general – including the class name, synonyms and scenes or situations where the class is likely to occur.

- **aeroplane**, airplane, plane, biplane, monoplane, aviator, bomber, hydroplane, airliner, aircraft, fighter, airport, hangar, jet, boeing, fuselage, wing, propellor, flying
- **bicycle**, bike, cycle, cyclist, pedal, tandem, saddle, wheel, cycling, ride, wheelie
- **bird**, birdie, birdwatching, nest, sea, aviary, birdcage, bird feeder, bird table,
- **boat** ship, barge, ferry, canoe, boating, craft, liner, cruise, sailing, rowing, watercraft, regatta, racing, marina, beach, water, canal, river, stream, lake, yacht,
- **bottle**, cork, wine, beer, champagne, ketchup, squash, soda, coke, lemonade, dinner, lunch, breakfast
- **bus**, omnibus, coach, shuttle, jitney, double-decker, motorbus, school bus, depot, terminal, station, terminus, passenger, route
- **car**, automobile, cruiser, motorcar, vehicle, hatchback, saloon, convertible, limousine, motor, race, traffic, trip, rally, city, street, road, lane, village, town, centre, shopping, downtown, suburban
- **cat**, feline, pussy, mew, kitten, tabby, tortoiseshell, ginger, stray
- **chair**, seat, rocker, rocking, deck, swivel, camp, chaise, office, studio, armchair, recliner, sitting, lounge, living room, sitting room
- **cow**, beef, heifer, moo, dairy, milk, milking, farm
- **dog**, hound, bark, kennel, heel, bitch, canine, puppy, hunter, collar, leash
- **horse**. gallop, jump, buck, equine, foal, cavalry, saddle, canter, buggy, mare, neigh, dressage, trial, racehorse, steeplechase, thoroughbred, cart, equestrian, paddock, stable, farrier
- **motorbike**, motorcycle, minibike, moped, dirt, pillion, biker, trials, motorcycling, motorcyclist, engine, motocross, scramble, sidecar, scooter, trail
- **person**, people, family, father, mother, brother, sister, aunt, uncle, grandmother, grandma, grandfather, grandpa, grandson, granddaughter, niece, nephew, cousin
- **sheep**, ram, fold, fleece, shear, baa, bleat, lamb, ewe, wool, flock
- **sofa**, chesterfield, settee, divan, couch, bolster
- **table**, dining, cafe, restaurant, kitchen, banquet, party, meal
- **potted plant**, pot plant, plant, patio, windowsill, window sill, yard, greenhouse, glass house, basket, cutting, pot, cooking, grow
- **train**, express, locomotive, freight, commuter, platform, subway, underground, steam, railway, railroad, rail, tube, underground, track, carriage, coach, metro, sleeper, railcar, buffet, cabin, level crossing
- **tv**, **monitor**, television, plasma, flatscreen, flat screen, lcd, crt, watching, dvd, desktop, computer, computer monitor, PC, console, game

and Everingham 2007), or being impossible to annotate correctly and completely with confidence.

One small bias was discovered in the VOC2007 dataset due to the image collection procedure – flickr returns query results ranked by "recency" such that if a given query is satisfied by many images, more recent images are returned first. Since the images were collected in January 2007, this led to an above-average number of Christmas/winter images containing, for example, large numbers of Christmas trees. To avoid such bias in VOC2008[5], images have been retrieved using queries comprising a random date in addition to keywords.

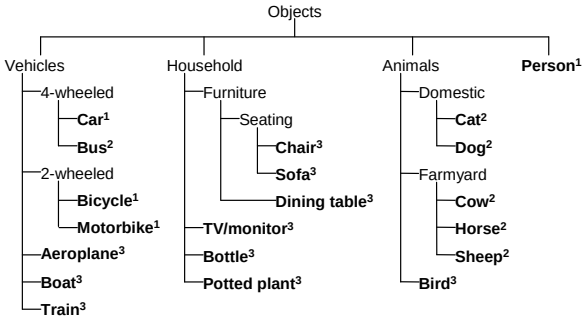


Fig. 2 VOC2007 Classes. Leaf nodes correspond to the 20 classes. The year of inclusion of each class in the challenge is indicated by superscripts: 2005[1], 2006[2], 2007[3]. The classes can be considered in a notional taxonomy, with successive challenges adding new branches (increasing the domain) and leaves (increasing detail).

3.2 Choice of Classes

Fig. 2 shows the 20 classes selected for annotation in the VOC2007 dataset. As shown, the classes can be considered in a taxonomy with four main branches – vehicles, animals, household objects and people[6]. The figure also shows the year of the challenge in which a particular class was included. In the original VOC2005 challenge (Everingham et al 2006a), which used existing annotated datasets, four classes were annotated (car, motorbike, bicycle and person). This number was increased to 10 in VOC2006, and 20 in VOC2007.

Over successive challenges the set of classes has been expanded in two ways: First, finer-grain "sub-classes" have been added e.g. "bus". The choice of sub-classes has been motivated by (i) increasing the "semantic" specificity of the output required of systems, for example recognising different types of vehicle e.g. car/motorbike (which may not be visually similar); (ii) increasing the difficulty of the discrimination task by inclusion of objects which might be considered visually similar e.g. "cat" vs. "dog". Second, additional branches of the notional taxonomy have been added e.g. "animals" (VOC2006) and "household objects" (VOC2007). The motivations are twofold: (i) increasing the domain of the challenge in terms of the semantic range of objects covered; (ii) encouraging research on object classes not widely addressed because of visual properties which are challenging for current methods, e.g. animals which might be considered to lack highly distinctive parts (c.f. car wheels), and chairs which are defined functionally, rather than visually, and also tend to be highly occluded in the dataset.

The choice of object classes, which can be considered a sub-tree of a taxonomy defined in terms of both semantic and visual similarity, also supports research in two areas which show promise in solving the scaling of object recognition to many thousands of classes: (i) exploiting visual properties common to classes e.g. vehicle wheels, for example in the form of "feature sharing" (Torralba et al 2007); (ii) exploiting external semantic information about the relations between object classes e.g. WordNet (Fellbaum 1998), for example by learning a hierarchy of classifiers (Marszalek and Schmid 2007). The availability of a class hierarchy may also prove essential in future evaluation efforts if the number of classes increases to the extent that there is implicit ambiguity in the classes, allowing individual objects to be annotated at different levels of the hierarchy e.g. hatchback/car/vehicle. We return to this point in Sect. 7.3.

3.3 Annotated Attributes

In order to evaluate the classification and detection challenges, the image annotation includes the following attributes for every object in the target set of object classes:

- **class:** one of: aeroplane, bird, bicycle, boat, bottle, bus, car, cat, chair, cow, dining table, dog, horse, motorbike, person, potted plant, sheep, sofa, train, tv/monitor.
- **bounding box:** an axis-aligned bounding box surrounding the extent of the object visible in the image.

The choice of an axis aligned bounding-box for the annotation is a compromise: for some object classes it fits quite well (e.g. to a horizontal bus or train) with only a small proportion of non-class pixels; however, for other classes it can be a poor fit either because they are not box shaped (e.g. a person with their arms outstretched, a chair) or/and because they are not axis-aligned (e.g. an aeroplane taking off). The advantage though is that they are relatively quick to annotate. We return to this point when discussing pixel level annotation in Sect. 3.6.1.

In addition, since VOC2006, further annotations were introduced which could be used during training but which were not required for evaluation:

[5] `http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2008/`

[6] These branches are also found in the Caltech 256 (Griffin et al 2007) taxonomy as transportation, animal, household & everyday, and human – though the Caltech 256 taxonomy has many other branches.

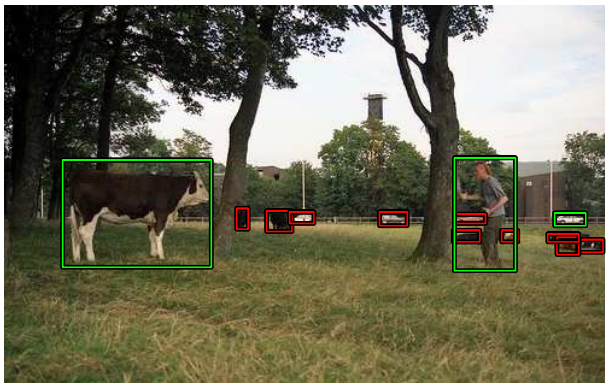


Fig. 3 Example of the "difficult" annotation. Objects shown in red have been marked difficult, and are excluded from the evaluation. Note that the judgement of difficulty is not solely by object size – the distant car on the right of the image is included in the evaluation.

- **viewpoint:** one of: front, rear, left, right, unspecified. This annotation supports methods which treat different viewpoints differently during training, such as using separate detectors for each viewpoint.
- **truncation:** an object is said to be "truncated" when the bounding box in the image does not correspond to the full extent of the object. This may occur for two reasons: (a) the object extends outside the image e.g. an image of a person from the waist up; (b) the boundary of the object is occluded e.g. a person standing behind a wall. The aim of including this annotation was to support recognition methods which require images of an *entire* object as training data, for example assuming that the bounding boxes of the objects can be aligned.

For the VOC2008 challenge, objects are additionally annotated as "occluded" if a high level of occlusion is present. This overcomes a limitation of the VOC2007 dataset that "clean" training examples without occlusion cannot automatically be identified from the available annotation.

- **difficult:** labels objects which are particularly difficult to detect due to small size, illumination, image quality or the need to use significant contextual information. In the challenge evaluation, such objects are discarded, although no penalty is incurred for detecting them. The aim of this annotation is to maintain a reasonable level of difficulty while not contaminating the evaluation with many near-unrecognisable examples.

Fig. 3 shows an example of the "difficult" annotation. The criteria used to judge an object difficult included confidence in the class label e.g. is it certain that all the animals in Fig. 3 are cows? (sometimes we see sheep in the same field), object size, level of occlusion, imaging factors e.g. motion blur, and requirement for significant context to enable recognition. Note that by marking difficult examples, rather than discarding them, the data should remain useful as methods able to cope with such examples are developed. Furthermore, as noted, any current methods able to detect difficult objects are not penalised for doing so.

3.4 Image Annotation Procedure

The VOC2007 annotation procedure was designed to be:

- **consistent**, so that the annotation of the images is consistent, in terms of the definition of the classes, how bounding boxes are placed, and how viewpoints and truncation are defined.
- **accurate**, so that there are as few annotation errors as possible,
- **exhaustive**, so that all object instances are labelled.

Consistency was achieved by having all annotation take place at a single annotation "party" at the University of Leeds, following a set of annotation guidelines which were discussed in detail with the annotators. The guidelines covered aspects including: what to label; how to label pose and bounding box; how to treat occlusion; acceptable image quality; how to label clothing/mud/snow, transparency, mirrors, and pictures. The full guidelines (Winn and Everingham 2007) are available on the WWW. In addition, during the annotation process, annotators were periodically observed to ensure that the guidelines were being followed. Several current annotation projects rely on untrained annotators or have annotators geographically distributed e.g. LabelMe (Russell et al 2008), or even ignorant of their task e.g. the ESP Game (von Ahn and Dabbish 2004). It is very difficult to maintain consistency of annotation in these circumstances, unlike when all annotators are trained, monitored and co-located.

Following the annotation party, the accuracy of each annotation was checked by one of the organisers, including checking for omitted objects to ensure exhaustive labelling. To date, only one error has been reported on the VOC2007 dataset, which was a viewpoint marked as unspecified rather than frontal. During the checking process, the "difficult" annotation was applied to objects judged as difficult to recognise. As checking the annotation is an extremely time-consuming process, for VOC2008 this has been incorporated into the annotation party, with each image checked for completeness and each object checked for accuracy, by one of the annotators. As in previous years, the "difficult" annotation was applied by one of the organisers to ensure consistency. We return to the question of the expense, in terms of person hours, of annotation and checking, in Sect. 7.3.

3.5 Dataset Statistics

Table 2 summarises the statistics of the VOC2007 dataset. For the purposes of the challenge, the data is divided

Table 2 Statistics of the VOC2007 dataset. The data is divided into two main subsets: training/validation data (`trainval`), and test data (`test`), with the `trainval` data further divided into suggested training (`train`) and validation (`val`) sets. For each subset and class, the number of images (containing at least one object of the corresponding class) and number of object instances are shown. Note that because images may contain objects of several classes, the totals shown in the image columns are not simply the sum of the corresponding column.

	`train`		`val`		`trainval`		`test`	
	img	**obj**	**img**	**obj**	**img**	**obj**	**img**	**obj**
Aeroplane	112	151	126	155	238	306	204	285
Bicycle	116	176	127	177	243	353	239	337
Bird	180	243	150	243	330	486	282	459
Boat	81	140	100	150	181	290	172	263
Bottle	139	253	105	252	244	505	212	469
Bus	97	115	89	114	186	229	174	213
Car	376	625	337	625	713	1,250	721	1,201
Cat	163	186	174	190	337	376	322	358
Chair	224	400	221	398	445	798	417	756
Cow	69	136	72	123	141	259	127	244
Dining table	97	103	103	112	200	215	190	206
Dog	203	253	218	257	421	510	418	489
Horse	139	182	148	180	287	362	274	348
Motorbike	120	167	125	172	245	339	222	325
Person	1,025	2,358	983	2,332	2,008	4,690	2,007	4,528
Potted plant	133	248	112	266	245	514	224	480
Sheep	48	130	48	127	96	257	97	242
Sofa	111	124	118	124	229	248	223	239
Train	127	145	134	152	261	297	259	282
Tv/monitor	128	166	128	158	256	324	229	308
Total	2,501	6,301	2,510	6,307	5,011	12,608	4,952	12,032

into two main subsets: training/validation data (`trainval`), and test data (`test`). For participants' convenience, the `trainval` data is further divided into suggested training (`train`) and validation (`val`) sets, however participants are free to use any data in the `trainval` set for training, for example if a given method does not require a separate validation set. The total number of annotated images is 9,963, roughly double the 5,304 images annotated for VOC2006. The number of annotated objects similarly rose from 9,507 to 24,640. Since the number of classes doubled from 10 to 20, the average number of objects of each class increased only slightly from 951 to 1,232, dominated by a quadrupling of the number of annotated people.

Fig. 4 shows a histogram of the number of images and objects in the entire dataset for each class. Note that these counts are shown on a log scale. The "person" class is by far the most frequent, with 9,218 object instances vs. 421 (dining table) to 2,421 (car) for the other classes. This is a natural consequence of requiring each image to be completely annotated – most flickr images can be characterised as "snapshots" e.g. family holidays, birthdays, parties, etc. and so many objects appear only "incidentally" in images where people are the subject of the photograph.

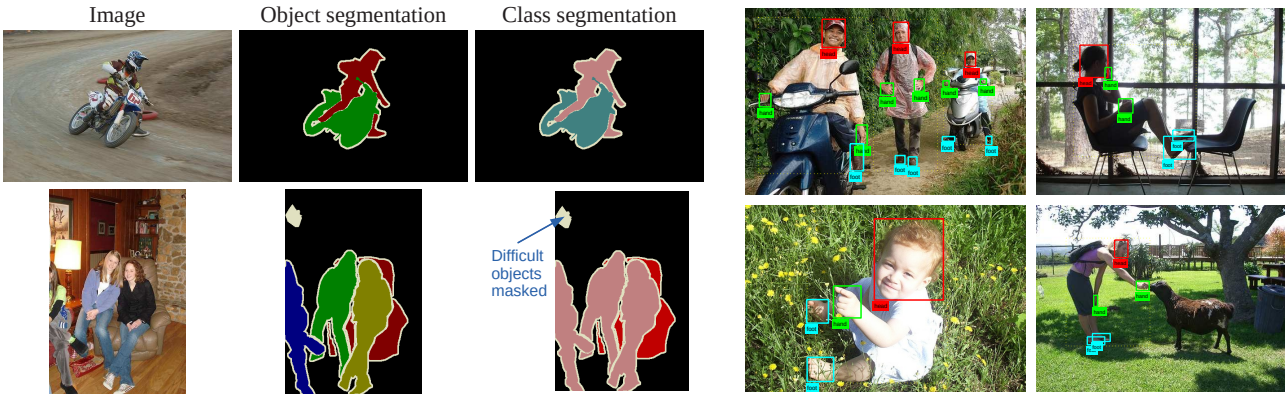
While the properties of objects in the dataset such as size and location in the image can be considered representative of flickr as a whole, the same cannot be said about the frequency of occurrence of each object class. In order to provide a reasonable minimum number of images/objects per class to participants, both for training and evaluation, certain minority classes e.g. "sheep" were targeted toward the end of the annotation party to increase their numbers – annotators were instructed to discard all images not containing one of the minority classes. Examples of certain classes e.g. "sheep" and "bus" proved difficult to collect, due either to lack of relevant keyword annotation by flickr users, or lack of photographs containing these classes.

3.6 Taster Competitions

Annotation was also provided for the newly introduced *segmentation* and *person layout* taster competitions. The idea behind these competitions is to allow systems to demonstrate a more detailed understanding of the image, such that objects can be localised down to the pixel level, or an object's parts (e.g. a person's head, hands and feet) can be localised within the object. As for the main competitions, the emphasis was on consistent, accurate and exhaustive annotation.

3.6.1 Segmentation

For the segmentation competition, a subset of images from each of the main datasets was annotated with pixel-level segmentations of the visible region of all contained objects. These segmentations act as a refinement of the bounding



(a) Segmentation taster (b) Person layout taster

Fig. 5 Example images and annotation for the taster competitions. (a) Segmentation taster annotation showing object and class segmentation. Border regions are marked with the "void" label indicating that they may be object or background. Difficult objects are excluded by masking with the 'void' label. (b) Person Layout taster annotation showing bounding boxes for head, hands and feet.

Table 3 Statistics of the VOC2007 segmentation dataset. The data is divided into two main subsets: training/validation data (`trainval`), and test data (`test`), with the `trainval` data further divided into suggested training (`train`) and validation (`val`) sets. For each subset and class, the number of images (containing at least one object of the corresponding class) and number of object instances are shown. Note that because images may contain objects of several classes, the totals shown in the image columns are not simply the sum of the corresponding column. All objects in each image are segmented, with every pixel of the image being labelled as one of the object classes, "background" (not one of the annotated classes) or "void" (uncertain i.e. near object boundary).

	`train`		`val`		`trainval`		`test`	
	img	**obj**	**img**	**obj**	**img**	**obj**	**img**	**obj**
Aeroplane	12	17	13	16	25	33	15	15
Bicycle	11	16	10	16	21	32	11	15
Bird	13	15	13	20	26	35	12	15
Boat	11	15	9	29	20	44	13	16
Bottle	17	30	13	28	30	58	13	20
Bus	14	16	11	15	25	31	12	17
Car	14	34	17	36	31	70	24	58
Cat	15	15	15	18	30	33	14	17
Chair	26	52	20	48	46	100	21	49
Cow	11	27	10	16	21	43	10	26
Diningtable	14	15	17	17	31	32	14	15
Dog	17	20	14	19	31	39	13	18
Horse	15	18	17	19	32	37	11	16
Motorbike	11	15	15	16	26	31	13	19
Person	92	194	79	154	171	348	92	179
Pottedplant	17	33	17	45	34	78	11	25
Sheep	8	41	13	22	21	63	10	27
Sofa	17	22	13	15	30	37	15	16
Train	8	14	15	17	23	31	16	17
Tvmonitor	20	24	13	16	33	40	17	27
Total	209	633	213	582	422	1,215	210	607

box, giving more precise shape and localisation information. In deciding how to provide pixel annotation, it was necessary to consider the trade-off between accuracy and annotation time: providing pixel-perfect annotation is extremely time intensive. To give high accuracy but to keep the annotation time short enough to provide a large image set, a border area of 5 pixels width was allowed around each object where the pixels were labelled neither object nor background (these were marked "void" in the data, see Fig. 5a). Annotators were also provided with detailed guidelines to ensure consistent segmentation (Winn and Everingham 2007). In keeping with the main competitions, difficult examples of objects were removed from both training and test sets by masking these objects with the "void" label.

The *object* segmentations, where each pixel is labelled with the identifier of a particular object, were used to create *class* segmentations (see Fig. 5a for examples) where each pixel is assigned a class label. These were provided

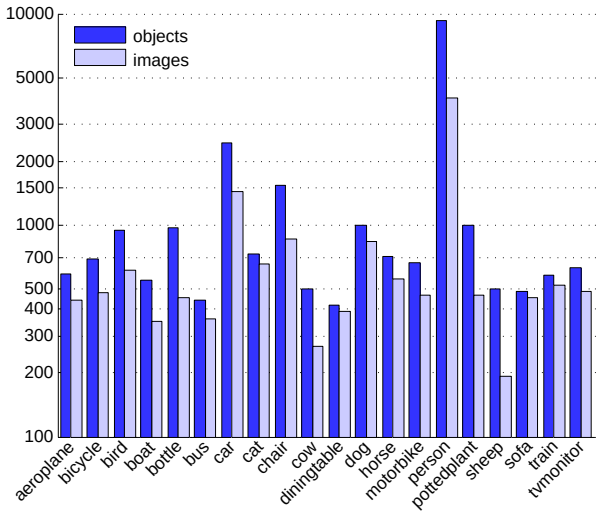


Fig. 4 Summary of the entire VOC2007 dataset. Histogram by class of the number of objects and images containing at least one object of the corresponding class. Note the log scale.

to encourage participation from class-based methods, which output a class label per pixel but which do not output an object identifier, e.g. do not segment adjacent objects of the same class. Participants' results were submitted in the form of class segmentations, where the aim is to predict the correct class label for every pixel not labelled in the ground truth as "void".

Table 3 summarises the statistics of the segmentation dataset. In total, 422 images containing 1,215 segmented objects were provided in the combined training/validation set. The test set contained 210 images and 607 objects.

3.6.2 Person layout

For the person layout competition, a subset of "person" objects in each of the main datasets was annotated with information about the 2-D pose or "layout" of the person. For each person, three types of "part" were annotated with bounding boxes: the head, hands, and feet, see Fig. 5b. These parts were chosen to give a good approximation of the overall pose of a person, and because they can be annotated with relative speed and accuracy compared to e.g. annotation of a "skeleton" structure where uncertainty in the position of the limbs and joints is hard to avoid. Annotators selected images to annotate which were of sufficient size such that there was no uncertainty in the position of the parts, and where the head and at least one other part were visible – no other criteria were used to "filter" suitable images. Fig. 5b shows some example images, including partial occlusion (upper-left), challenging lighting (upper-right), and "non-standard" pose (lower-left). In total, the training/validation set contained 439 annotated people in 322 images, and the test set 441 annotated people in 441 images.

4 Submission and Evaluation

The submission and evaluation procedures for the VOC2007 challenge competitions were designed to be fair, to prevent over-fitting, and to demonstrate clearly the differences in accuracy between different methods.

4.1 Submission of Results

The running of the VOC2007 challenge consisted of two phases: At the start of the challenge, participants were issued a development kit comprising training/validation images with annotation, and MATLAB[7] software to access the annotation (stored in an XML format compatible with LabelMe (Russell et al 2008)), to compute the evaluation measures, and including simple baseline implementations for each competition. In the second phase, *un-annotated* test images were distributed. Participants were then required to run their methods on the test data and submit results as defined in Sect. 4.2. The test data was available for approximately three months before submission of results – this allowed substantial time for processing, and aimed to not penalise computationally expensive methods, or groups with access to only limited computational resources.

Withholding the annotation of the test data until completion of the challenge played a significant part in preventing over-fitting of the parameters of classification or detection methods. In the VOC2005 challenge, test annotation was released and this led to some "optimistic" reported results, where a number of parameter settings had been run on the test set, and only the best reported. This danger emerges in any evaluation initiative where ground truth is publicly available. Because the test data is in the form of images, it is also theoretically possible for participants to hand-label the test data, or "eyeball" test results – this is in contrast to e.g. machine learning benchmarks where the test data may be sufficiently "abstract" such that it cannot easily be labelled by a non-specialist. We rely on the participants' honesty, and the limited time available between release of the test data and submission of results, to minimise the possibility of manual labelling. The possibility could be avoided by requiring participants to submit code for their methods, and never release the test images. However, this makes the evaluation task difficult for both participants and organisers, since methods may use a mixture of MATLAB/C code, proprietary libraries, require significant computational resources, etc. It is worth noting, however, that results submitted to the VOC challenge, rather than afterward using the released annotation data, might appropriately be accorded higher status since participants have limited opportunity to experiment with the test data.

[7] MATLAB® is a registered trademark of The MathWorks, Inc.

In addition to withholding the test data annotation, it was also required that participants submit only a *single* result per method, such that the organisers were not asked to choose the best result for them. Participants were not required to provide classification or detection results for all 20 classes, to encourage participation from groups having particular expertise in e.g. person or vehicle detection.

4.2 Evaluation of Results

Evaluation of results on multi-class datasets such as VOC2007 poses several problems: (i) for the classification task, images contain instances of multiple classes, so a "forced choice" paradigm such as that adopted by Caltech 256 (Griffin et al 2007) – "which one of *m* classes does this image contain?" – cannot be used; (ii) the prior distribution over classes is significantly nonuniform so a simple *accuracy* measure (percentage of correctly classified examples) is not appropriate. This is particularly salient in the detection task, where sliding window methods will encounter many thousands of negative (non-class) examples for every positive example. In the absence of information about the *cost* or risk of misclassifications, it is necessary to evaluate the trade-off between different types of classification error; (iii) evaluation measures need to be algorithm-independent, for example in the detection task participants have adopted a variety of methods e.g. sliding window classification, segmentation-based, constellation models, etc. This prevents the use of some previous evaluation measures such as the Detection Error Tradeoff (DET) commonly used for evaluating pedestrian detectors (Dalal and Triggs 2005), since this is applicable only to sliding window methods constrained to a specified window extraction scheme, and to data with cropped positive test examples.

Both the classification and detection tasks were evaluated as a set of 20 independent two-class tasks: e.g. for classification "is there a car in the image?", and for detection "where are the cars in the image (if any)?". A separate "score" is computed for each of the classes. For the classification task, participants submitted results in the form of a confidence level for each image and for each class, with larger values indicating greater confidence that the image contains the object of interest. For the detection task, participants submitted a bounding box for each detection, with a confidence level for each bounding box. The provision of a confidence level allows results to be ranked such that the trade-off between false positives and false negatives can be evaluated, without defining arbitrary costs on each type of classification error.

Average Precision (AP). For the VOC2007 challenge, the interpolated average precision (Salton and Mcgill 1986) was used to evaluate both classification and detection.

For a given task and class, the precision/recall curve is computed from a method's ranked output. Recall is defined as the proportion of all positive examples ranked above a given rank. Precision is the proportion of all examples above that rank which are from the positive class. The AP summarises the shape of the precision/recall curve, and is defined as the mean precision at a set of eleven equally spaced recall levels $[0, 0.1, \ldots, 1]$:

$$AP = \frac{1}{11} \sum_{r \in \{0, 0.1, \ldots, 1\}} p_{interp}(r) \tag{1}$$

The precision at each recall level r is *interpolated* by taking the maximum precision measured for a method for which the corresponding recall exceeds r:

$$p_{interp}(r) = \max_{\tilde{r}:\tilde{r} \geq r} p(\tilde{r}) \tag{2}$$

where $p(\tilde{r})$ is the measured precision at recall $\tilde{r}$.

The intention in interpolating the precision/recall curve in this way is to reduce the impact of the "wiggles" in the precision/recall curve, caused by small variations in the ranking of examples. It should be noted that to obtain a high score, a method must have precision at all levels of recall – this penalises methods which retrieve only a subset of examples with high precision (e.g. side views of cars).

The use of precision/recall and AP replaced the "area under curve" (AUC) measure of the ROC curve used in VOC2006 for the classification task. This change was made to improve the sensitivity of the metric (in VOC2006 many methods were achieving greater than 95% AUC), to improve interpretability (especially for image retrieval applications), to give increased visibility to performance at low recall, and to unify the evaluation of the two main competitions. A comparison of the two measures on VOC2006 showed that the ranking of participants was generally in agreement but that the AP measure highlighted differences between methods to a greater extent.

Bounding box evaluation. As noted, for the detection task, participants submitted a list of bounding boxes with associated confidence (rank). Detections were assigned to ground truth objects and judged to be true/false positives by measuring bounding box overlap. To be considered a correct detection, the area of overlap a_o between the predicted bounding box B_p and ground truth bounding box B_{gt} must exceed 0.5 (50%) by the formula

$$a_o = \frac{area(B_p \cap B_{gt})}{area(B_p \cup B_{gt})} \tag{3}$$

where $B_p \cap B_{gt}$ denotes the intersection of the predicted and ground truth bounding boxes and $B_p \cup B_{gt}$ their union.

The threshold of 50% was set deliberately low to account for inaccuracies in bounding boxes in the ground truth data, for example defining the bounding box for a highly non-convex object, e.g. a person with arms and legs spread, is

somewhat subjective. Sect. 6.2.3 evaluates the effect of this threshold on the measured average precision. We return to the question of the suitability of bounding box annotation in Sect. 7.3.

Detections output by a method were assigned to ground truth objects satisfying the overlap criterion in order ranked by the (decreasing) confidence output. Multiple detections of the same object in an image were considered false detections e.g. 5 detections of a single object counted as 1 correct detection and 4 false detections – it was the responsibility of the participant's system to filter multiple detections from its output.

4.2.1 Evaluation of the segmentation taster

A common measure used to evaluate segmentation methods is the percentage of pixels correctly labelled. For the VOC2007 segmentation taster, this measure was used per class by considering only pixels labelled with that class in the ground truth annotation. Reporting a per-class accuracy in this way allowed participants to enter segmentation methods which handled only a subset of the classes. However, this evaluation scheme can be misleading, for example, labelling *all* pixels "car" leads to a perfect score on the car class (though not the other classes). Biases in different methods can hence lead to misleading high or low accuracies on individual classes. To rectify this problem, the VOC2008 segmentation challenge will be assessed on a modified per-class measure based on the intersection of the inferred segmentation and the ground truth, divided by the union:

$$\text{seg. accuracy} = \frac{\text{true pos.}}{\text{true pos.} + \text{false pos.} + \text{false neg.}} \quad (4)$$

Pixels marked "void" in the ground truth are excluded from this measure. Compared to VOC2007, the measure penalises methods which have high false positive rates (i.e. that incorrectly mark non-class pixels as belonging to the target class). The per-class measure should hence give a more interpretable evaluation of the performance of individual methods.

4.2.2 Evaluation of the person layout taster

The "person layout" taster was treated as an extended detection task. Methods were evaluated using the same AP measure used for the main detection competition. The criterion for a correct detection, however, was extended to require correct prediction of (i) the set of visible parts (head/hands/feet); (ii) correct bounding boxes for all parts, using the standard overlap threshold of 50%.

As reported in Sect. 6.4 this evaluation criterion proved extremely challenging. In the VOC2008 challenge, the evaluation has been relaxed by providing person bounding boxes for the test data (disjoint from the main challenge test set), so that methods are not required to complete the detection part of the task, but only estimate part identity and location.

5 Methods

Table 4 summarises the participation in the VOC2007 challenge. A total of 16 institutions submitted results (c.f. 16 in 2006 and 9 in 2005). Taking into account multiple groups in an institution and multiple methods per group, there were a total of 28 methods submitted (c.f. 25 in 2006, 13 in 2005).

5.1 Classification Methods

There were 17 entries for the classification task in 2007, compared to 14 in 2006 and 9 in 2005.

Many of the submissions used variations on the basic bag-of-visual-words method (Csurka et al (2004); Sivic and Zisserman (2003)) that was so successful in VOC2006, see Zhang et al (2007): local features are computed (for example SIFT descriptors); vector quantised (often by using *k*-means) into a visual vocabulary or codebook; and each image is then represented by a histogram of how often the local features are assigned to each visual word. The representation is known as bag-of-visual-words in analogy with the bag-of-words (BOW) text representation where the frequency, but not the position, of words is used to represent text documents. It is also known as bag-of-keypoints or bag-of-features. The classifier is typically a support vector machine (SVM) with χ^2 or Earth Mover's Distance (EMD) kernel.

Within this approach, submissions varied tremendously in the features used: both their type and their density. Sparse local features were detected using the Harris interest point operator and/or the SIFT detector (Lowe 2004), and then represented by the SIFT descriptor. There was some attention to exploring different colour spaces (such as HSI) in the detection for greater immunity to photometric effects such as shadows (*PRIP-UvA*). Others (e.g. *INRIA_Larlus*) computed descriptors on a dense grid, and one submission (*MPI*) combined both sparse and dense descriptors. In addition to SIFT, other descriptors included local colour, pairs of adjacent segments (PAS) (Ferrari et al 2008), and Sobel edge histograms.

The BOW representation was still very common, where spatial information, such as the position of the descriptors is disregarded. However, several participants provided additional representations (channels) for each image where as well as the BOW, spatial information was included by various tilings of the image (*INRIA_Genetic*, *INRIA_Flat*), or using a spatial pyramid (*TKK*).

While most submissions used a kernel SVM as the classifier (with kernels including χ^2 and EMD), *XRCE* used lo-

Table 4 Participation in the VOC2007 challenge. Each method is assigned an abbreviation used in the text, and identified as a classification method (Cls) or detection method (Det). The contributors to each method are listed with references to publications describing the method, where available.

Abbreviation	Cls	Det	Contributors	References
Darmstadt	–	•	Mario Fritz and Bernt Schiele, TU Darmstadt	Fritz and Schiele (2008)
INRIA_Flat	•	–	Marcin Marszalek, Cordelia Schmid, Hedi Harzallah and Joost Van-de-weijer, INRIA Rhone-Alpes	Zhang et al (2007); van de Weijer and Schmid (2006); Ferrari et al (2008)
INRIA_Genetic	•	–		
INRIA_Larlus	•	–	Diane Larlus and Frederic Jurie, INRIA Rhones-Alpes	–
INRIA_Normal	–	•	Hedi Harzallah, Cordelia Schmid, Marcin Marszalek, Vittorio Ferrari, Y-Lan Boureau, Jean Ponce and Frederic Jurie, INRIA Rhone-Alpes	Ferrari et al (2008); van de Weijer and Schmid (2006); Zhang et al (2007)
INRIA_PlusClass	–	•		
IRISA	–	•	Ivan Laptev, IRISA/INRIA Rennes and Evgeniy Tarassov, TT-Solutions	Laptev (2006)
MPI_BOW	•	–	Christoph Lampert and Matthew Blaschko, MPI Tuebingen	Lampert et al (2008)
MPI_Center	–	•		
MPI_ESSOL	–	•		
Oxford	–	•	Ondrej Chum and Andrew Zisserman, University of Oxford	Chum and Zisserman (2007)
PRIPUVA	•	–	Julian Stottinger and Allan Hanbury, Vienna University of Technology; Nicu Sebe and Theo Gevers, University of Amsterdam	Stoettinger et al (2007)
QMUL_HSLS	•	–	Jianguo Zhang, Queen Mary University of London	Zhang et al (2007)
QMUL_LSPCH	•	–		
TKK	•	•	Ville Viitaniemi and Jorma Laaksonon, Helsinki University of Technology	Viitaniemi and Laaksonen (2008)
ToshCam_rdf	•	–	Jamie Shotton, Toshiba Corporate R&D Center, Japan & Matthew Johnson, University of Cambridge	–
ToshCam_svm	•	–		
Tsinghua	•	–	Dong Wang, Xiaobing Liu, Cailiang Liu, Zhang Bo and Jianmin Li, Tsinghua University	Wang et al (2006); Liu et al (2007)
UoCTTI	–	•	Pedro Felzenszwalb, University of Chicago; David McAllester and Deva Ramanan, Toyota Technological Institute, Chicago	Felzenszwalb et al (2008)
UVA_Bigrams	•	–	Koen van de Sande, Jan van Gemert and Jasper Uijlings, University of Amsterdam	van de Sande et al (2008); van Gemert et al (2006); Geusebroek (2006); Snoek et al (2006, 2005)
UVA_FuseAll	•	–		
UVA_MCIP	•	–		
UVA_SFS	•	–		
UVA_WGT	•	–		
XRCE	•	–	Florent Perronnin, Yan Liu and Gabriela Csurka, Xerox Research Centre Europe	Perronnin and Dance (2007)

Table 5 Classification results. For each object class and submission, the AP measure (%) is shown. Bold entries in each column denote the maximum AP for the corresponding class. Italic entries denote the results ranked second or third.

	aeroplane	bicycle	bird	boat	bottle	bus	car	cat	chair	cow	diningtable	dog	horse	motorbike	person	pottedplant	sheep	sofa	train	tvmonitor
INRIA_Flat	*74.8*	*62.5*	*51.2*	*69.4*	*29.2*	*60.4*	*76.3*	*57.6*	*53.1*	*41.1*	*54.0*	*42.8*	*76.5*	*62.3*	*84.5*	*35.3*	*41.3*	*50.1*	*77.6*	*49.3*
INRIA_Genetic	**77.5**	**63.6**	**56.1**	**71.9**	**33.1**	**60.6**	**78.0**	**58.8**	**53.5**	**42.6**	**54.9**	**45.8**	**77.5**	**64.0**	**85.9**	**36.3**	**44.7**	*50.6*	**79.2**	**53.2**
INRIA_Larlus	62.6	54.0	32.8	47.5	17.8	46.4	69.6	44.2	44.6	26.0	38.1	34.0	66.0	55.1	77.2	13.1	29.1	36.7	62.7	43.3
MPI_BOW	58.9	46.0	31.3	59.0	16.9	40.5	67.2	40.2	44.3	28.3	31.9	34.4	63.6	53.5	75.7	22.3	26.6	35.4	60.6	40.6
PRIPUVA	48.6	20.9	21.3	17.2	6.4	14.2	45.0	31.4	27.4	12.3	14.3	23.7	30.1	13.3	62.0	10.0	12.4	13.3	26.7	26.2
QMUL_HSLS	70.6	54.8	35.7	64.5	27.8	51.1	71.4	54.0	46.6	36.6	34.4	39.9	71.5	55.4	80.6	15.8	35.8	41.5	73.1	45.5
QMUL_LSPCH	71.6	55.0	41.1	65.5	27.2	51.1	72.2	*55.1*	47.4	35.9	37.4	41.5	71.5	57.9	80.8	15.6	33.3	41.9	*76.5*	45.9
TKK	71.4	51.7	48.5	63.4	27.3	49.9	70.1	51.2	51.7	32.3	46.3	41.5	72.6	*60.2*	82.2	31.7	30.1	39.2	71.1	41.0
ToshCam_rdf	59.9	36.8	29.9	40.0	23.6	33.3	60.2	33.0	41.0	17.8	33.2	33.7	63.9	53.1	77.9	29.0	27.3	31.2	50.1	37.6
ToshCam_svm	54.0	27.1	30.3	35.6	17.0	22.3	58.0	34.6	38.0	19.0	27.5	32.4	48.0	40.7	78.1	23.4	21.8	28.0	45.5	31.8
Tsinghua	62.9	42.4	33.9	49.7	23.7	40.7	62.0	35.2	42.7	21.0	38.9	34.7	65.0	48.1	76.9	16.9	30.8	32.8	58.9	33.1
UVA_Bigrams	61.2	33.2	29.4	45.0	16.5	37.6	54.6	31.3	39.9	17.2	31.4	30.6	61.6	42.4	74.6	14.5	20.9	23.5	49.9	30.0
UVA_FuseAll	67.1	48.1	43.3	58.1	19.9	46.3	61.8	41.9	48.4	27.8	41.9	38.5	69.8	51.4	79.4	32.5	31.9	36.0	66.2	40.3
UVA_MCIP	66.5	47.9	41.0	58.0	16.8	44.0	61.2	40.5	48.5	27.8	41.7	37.1	66.4	50.1	78.6	31.2	32.3	31.9	66.6	40.3
UVA_SFS	66.3	49.7	43.5	60.7	18.8	44.9	64.8	41.9	46.8	24.9	42.3	33.9	71.5	53.4	80.4	29.7	31.2	31.8	67.4	43.5
UVA_WGT	59.7	33.7	34.9	44.5	22.2	32.9	55.9	36.3	36.8	20.6	25.2	34.7	65.1	40.1	74.2	26.4	26.9	25.1	50.7	29.7
XRCE	*72.3*	*57.5*	*53.2*	*68.9*	*28.5*	*57.5*	*75.4*	50.3	*52.2*	*39.0*	*46.8*	*45.3*	*75.7*	58.5	*84.0*	*32.6*	*39.7*	**50.9**	75.1	*49.5*

gistic regression with a Fisher kernel (Perronnin and Dance 2007), and *ToshCam* used a random forest classifier.

Where there was greatest diversity was in the methods for combining the multiple representations (channels). Some methods investigated "late fusion" where a classifier is trained on each channel independently, and then a second classifier combines the results. For example *TKK* used this approach, for details see Viitaniemi and Laaksonen (2008). *Tsinghua* combined the individual classifiers using RankBoost. *INRIA* entered two methods using the same channels, but differing in the manner in which they were combined: *INRIA_Flat* uses uniform weighting on each feature (following Zhang et al (2007)); *INRIA_Genetic* uses a different class-dependent weight for each feature, learnt from the validation data by a genetic algorithm search.

In 2006, several of the submissions tackled the classification task as detection – "there is a car here, so the image contains a car". This approach is perhaps more in line with human intuition about the task, in comparison to the "global" classification methods which establish the presence of an object without localising it in the image. However, in 2007 no submissions used this approach.

The VOC challenge invites submission of results from "off-the-shelf" systems or methods trained on data other than that provided for the challenge (see Sect. 2.1), to be evaluated separately from those using only the provided data. No results were submitted to VOC2007 in this category. This is disappointing, since it prevents answering the question as to how well current methods perform given unlimited training data, or more detailed annotation of training data. It is an open question how to encourage submission of results from e.g. commercial systems.

5.2 Detection Methods

There were 9 entries for the detection task in 2007, compared to 9 in 2006 and 5 in 2005. As for the classification task, all submitted methods were trained only on the provided training data.

The majority of the VOC2007 entries used a "sliding window" approach to the detection task or variants thereof. In the basic sliding window method a rectangular window of the image is taken, features are extracted from this window, and it is then classified as either containing an instance of a given class or not. This classifier is then run exhaustively over the image at varying location and scale. In order to deal with multiple nearby detections a "non-maximum suppression" stage is then usually applied. Prominent examples of this method include the Viola and Jones (2004) face detector and the Dalal and Triggs (2005) pedestrian detector.

The entries *Darmstadt*, *INRIA_Normal*, *INRIA_PlusClass* and *IRISA* were essentially sliding window methods, with the enhancements that *INRIA_PlusClass* also utilised the output of a whole image classifier, and that *IRISA* also trained separate detectors for person-on-X where X was horse, bicycle, or motorbike. Two variations on the sliding window method avoided dense sampling of the test image: The *Oxford* entry used interest point detection to select candidate windows, and then applied an SVM classifier; see Chum and Zisserman (2007) for details. The *MPI_ESSOL* entry (Lampert et al 2008) used a branch-and-bound scheme to efficiently maximise the classifier function (based on a BOW representation, or pyramid match kernel (Grauman and Darrell 2005) determined on a per-class basis at training time) over all possible windows.

The *UoCTTI* entry used a more complex variant of the sliding window method, see Felzenszwalb et al (2008) for details. It combines the outputs of a coarse window and sev-

eral higher-resolution part windows which can move relative to the coarse window; inference over location of the parts is performed for each coarse image window. Note that improved results are reported in Felzenszwalb et al (2008) relative to those in Table 6; these were achieved after the public release of the test set annotation.

The method proposed by *TKK* automatically segments an image to extract candidate bounding boxes and then classifies these bounding boxes, see Viitaniemi and Laaksonen (2008) for details. The *MPI_Center* entry was a baseline that returns exactly one object bounding box per image; the box is centred and is 51% of the total image area.

In previous VOC detection competitions there had been a greater diversity of methods used for the detection problem, see Everingham et al (2006b) for more details. For example in VOC2006 the *Cambridge* entry used a classifier to predict a class label at each pixel, and then computed contiguously segmented regions; the *TU Darmstadt* entry made use of the Implicit Shape Model (ISM) (Leibe et al 2004); and the *MIT_Fergus* entry used the "constellation" model (Fergus et al 2007).

6 Results

This section reports and discusses the results of the VOC2007 challenge. Full results including precision/recall curves for all classes, not all of which are shown here due to space constraints, can be found on the VOC2007 website (Everingham et al 2007).

6.1 Classification

This section reports and discusses the results of the classification task. A total of 17 methods were evaluated. All participants tackled all of the 20 classes. Table 5 lists the AP for all submitted methods and classes. For each class the method obtaining the greatest AP is identified in bold, and the methods with 2nd and 3rd greatest AP in italics. Precision/recall curves for a representative sample of classes are shown in Fig. 6. Results are shown ordered by decreasing maximum AP. The left column shows all results, while the right column shows the top five results by AP. The left column also shows the "chance" performance, obtained by a classifier outputting a random confidence value without examining the input image – the precision/recall curve and corresponding AP measure are not invariant to the proportion of positive images, resulting in varying chance performance across classes. We discuss this further in Sect. 6.1.3.

6.1.1 Classification results by method

Overall the *INRIA_Genetic* method stands out as the most successful method, obtaining the greatest AP in 19 of the 20 classes. The related *INRIA_Flat* method achieves very similar performance, with AP between the two methods differing by just 1–2% for most classes. As described in Sect. 5.1 these methods use the same set of heterogeneous image features, and differ only in the way that features are fused in a generalised radial basis function (RBF) kernel: *INRIA_Flat* uses uniform weighting on each feature, and *INRIA_Genetic* learns a different weight for each feature from the validation data. The *XRCE* method comes third in 17 of 20 classes and first in one. This method differs from the INRIA methods in using a Fisher kernel representation of the distribution of visual features within the image, and uses a smaller feature set and logistic regression cf. the kernel SVM classifier used by the INRIA methods.

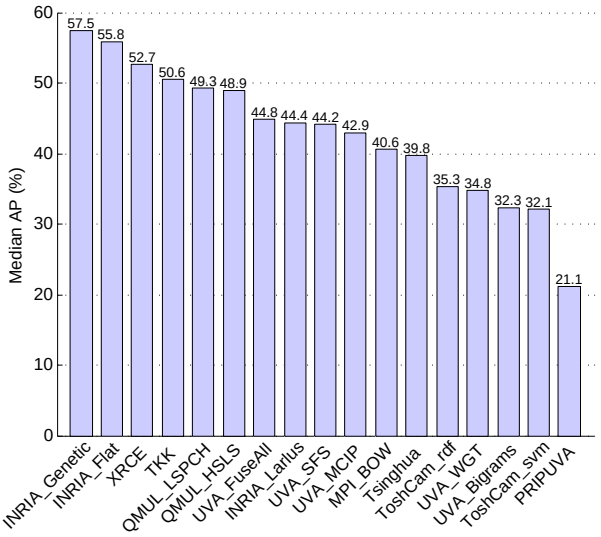


Fig. 7 Summary of the classification results by method. For each method the median AP over all classes is shown.

Fig. 7 summarises the performance of all methods, plotting the median AP for each method computed over all classes, and ordered by decreasing median AP. Despite the overall similarity in the features used, there is quite a wide range in accuracy of 32.1–57.5%, with one method (*PRIPUVA*) substantially lower at 21.1%. The high performing methods all combine multiple features (channels) though, and some (*INRIA_Genetic*, *INRIA_Flat*, *TKK*) include spatial information as well as BOW. Software for the feature descriptors used by the *UVA* methods (van de Sande et al 2008) has been made publicly available, and would form a reasonable state-of-the-art baseline for future challenges.

6.1.2 Statistical significance of results

A question often overlooked by the computer vision community when comparing results on a given dataset is whether the difference in performance of two methods is statistically

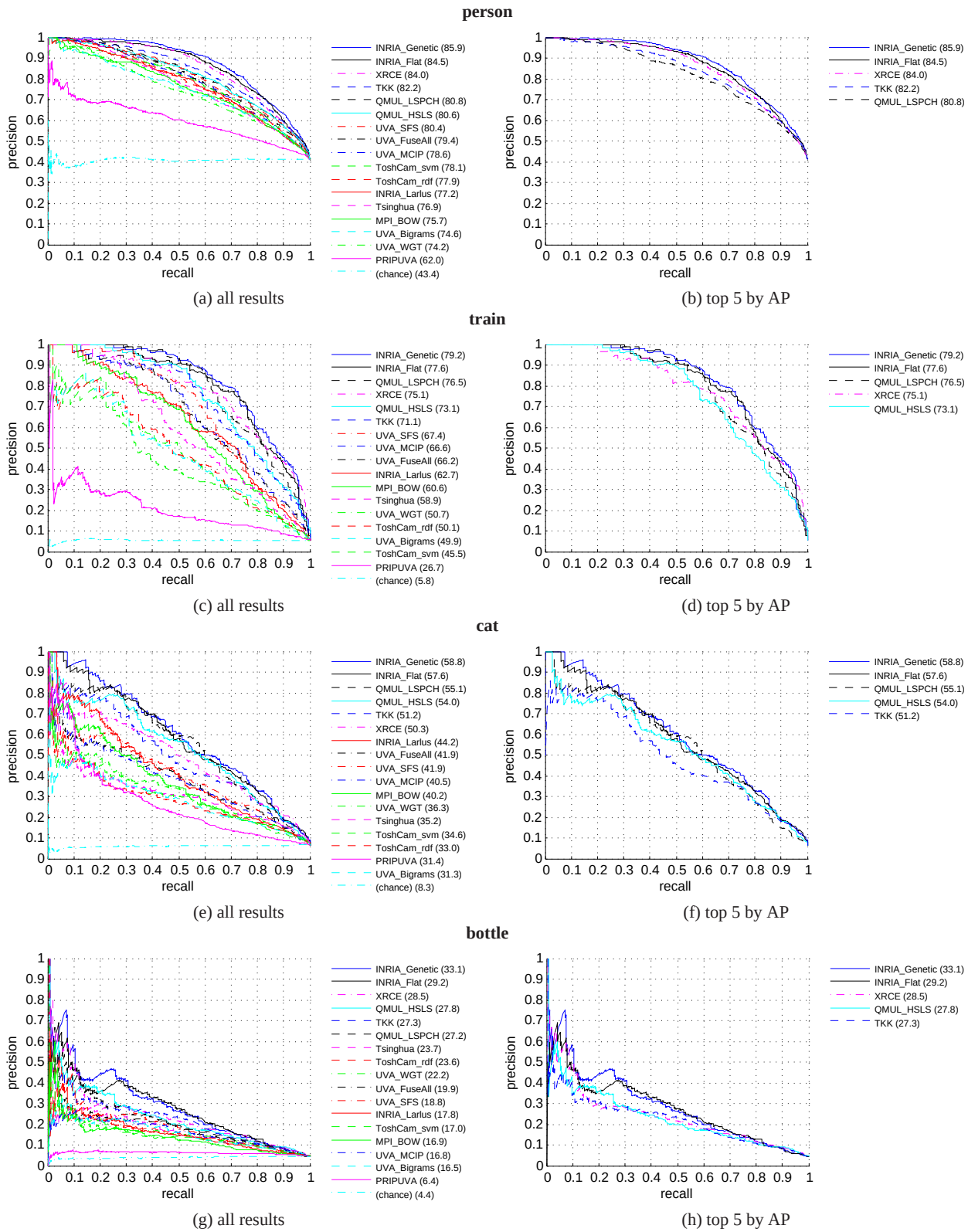


Fig. 6 Classification results. Precision/recall curves are shown for a representative sample of classes. The left column shows all results; the right shows the top 5 results by AP. The legend indicates the AP (%) obtained by the corresponding method.

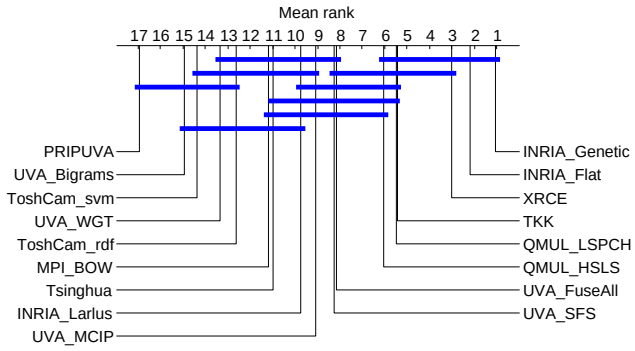


Fig. 8 Analysis of statistically significant differences in the classification results. The mean rank over all classes is plotted on the *x*-axis for each method. Methods which are not significantly different ($p = 0.05$), in terms of mean rank, are connected.

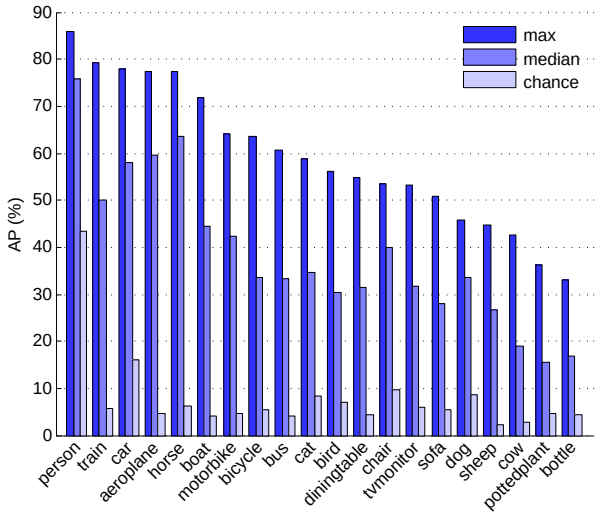


Fig. 9 Summary of classification results by class. For each class three values are shown: the maximum AP obtained by any method (max), the median AP over all methods (median) and the AP obtained by a random ranking of the images (chance).

significant. For the VOC challenge, we wanted to establish whether, for the given dataset and set of classes, one method can be considered significantly more accurate than another. Note that this question is different to that investigated e.g. in the Caltech 101 challenge, where multiple training/test folds are used to establish the variance of the measured accuracy. Whereas that approach measures robustness of a method to differing data, we wish to establish significance for the given, fixed, dataset. This is salient, for example, when a method may not involve a training phase, or to compare against a commercial system trained on proprietary data.

Little work has considered the comparison of multiple classifiers over multiple datasets. We analysed the results of the classification task using a method proposed by Demsar (2006), specifically using the Friedman test with Nemenyi *post hoc* analysis. This approach uses only comparisons between the *rank* of a method (the method achieving the greatest AP is assigned rank 1, the 2nd greatest AP rank 2, etc.), and thus requires no assumptions about the distribution of AP to be made. Each class is treated as a separate test, giving one rank measurement per method and class. The analysis then consists of two steps: (i) the null hypothesis is made that the methods are equivalent and so their ranks should be equal. The hypothesis is tested by the Friedman test (a non-parametric variant of ANOVA), which follows a χ^2 distribution; (ii) having rejected the null hypothesis the differences in ranks are analysed by the Nemenyi test (similar to the Tukey test for ANOVA). The difference between mean ranks (over classes) for a pair of methods follows a modified Studentised range statistic. For a confidence level of $p = 0.05$ and given the 17 methods tested over 20 classes, the "critical difference" is calculated as 4.9 – the difference in mean rank between a pair of methods must exceed 4.9 for the difference to be considered statistically significant.

Fig. 8 visualises the results of this analysis using the CD (critical difference) diagram proposed by Demsar (2006). The *x*-axis shows the mean rank over classes for each method. Methods are shown clockwise from right to left in decreasing (first to last) rank order. Groups of methods for which the difference in mean rank is not significant are connected by horizontal bars. As can be seen, there is substantial overlap between the groups, with no clear "clustering" into sets of equivalent methods. Of interest is that the differences between the first six ranked methods (*INRIA_Genetic*, *INRIA_Flat*, *XRCE*, *TKK*, *QMUL_LSPCH*, *QMUL_HSLS*) cannot be considered statistically significant.

A limitation of this analysis is that the relatively small number of observations (20 classes per method) limits the power of the test. Increasing the number of classes will make it more feasible to establish significant differences between methods in terms of their performance over a wide range of classes. As discussed in Sect. 7.1, we are also keen to highlight differences in the *approach* taken by methods, not solely their performance.

6.1.3 Classification results by class

Fig. 9 summarises the results obtained by object class, plotting for each class the maximum and median AP taken over all methods. Also shown is the "chance" performance – the AP obtained by a classifier outputting a random confidence value without examining the input image. Results are shown ordered by decreasing maximum AP. There is substantial variation in the maximum AP as a function of object class, from 33.1% (bottle) to 85.9% (person). The median AP varies from 15.6% (potted plant) to 75.7% (person). The median results can be seen to approximately follow the ranking of results by maximum AP, suggesting that the same classes proved difficult across methods, but individual differences

can be seen, for example the difference in maximum AP for the 2nd to 4th classes is very small such that the ordering is somewhat arbitrary. The high AP for the "person" class can be attributed in part to the high proportion of images in the dataset containing people – the chance AP for this class is 43.4%. However, as can be seen in Fig. 9, the results are substantially above chance for all classes.

While results on some classes e.g. person and train (Fig. 6a–d) are very promising, for all classes there is substantial room for improvement. For some classes e.g. bottle (Fig. 6g–h), the precision/recall curves show that the methods' precision drops greatly at moderate recall, and current methods would not give satisfactory results in the context of an image retrieval system. It is also worth noting that if one views the classification task as image retrieval, the evaluation is somewhat "benign" since the prior probability of an image containing an object of interest is still quite high, 2% for the least frequent class (sheep). We might expect that in a real world scenario, for example image-based web search, the prior probability for some classes would be much lower. We return to this point in Sect. 7.3.

6.1.4 What are the classification methods learning?

As noted, the quantitative evaluation of methods by AP gives a summary of a method's precision/recall trade-off. It is interesting to examine the success and failure modes of the methods to derive some insight into what current methods are learning, and what limitations might be addressed in development of future methods.

We first examined the kinds of images which methods found "easy" or "difficult". Five submitted methods were selected which represented somewhat different approaches rather than small variations (e.g. *INRIA_Genetic* vs. *INRIA_Flat*): *INRIA_Genetic*, *XRCE*, *TKK*, *QMUL_LSPCH* and *UVA_FuseAll*. Each test image was then assigned a rank by each method (using the method's confidence output for that image). An overall rank for the image was then assigned by taking the median over the ranks from the five selected methods. By looking at which images, containing the class of interest or not, are ranked first or last, we can gain insight into what properties of the images make recognition easy or difficult for current methods.

Fig. 10 shows ranked images for the "car" class. The first row shows the five positive images (containing cars) assigned the highest rank (1st–5th) – these can be considered images which are "easy" for current methods to recognise as containing cars. The second row shows the five positive images (containing cars) assigned the lowest rank – these are images for which current methods cannot easily establish the presence of a car. The third row shows the five negative images (not containing cars) assigned the highest rank – these are images which "confuse" current methods, which judge them highly likely to contain cars.

The high ranked positive images (Fig. 10a) include images where a single car dominates the image, in an uncluttered or "expected" background i.e. a road, and images where a number of cars are visible. The inclusion of images with multiple cars is perhaps surprising, but as discussed below, may be attributed to the reliance of current methods on "textural" properties of the image rather than spatial arrangement of parts. The low ranked positive images (Fig. 10b) are typical across all classes, showing the object of interest small in the image, poorly lit or heavily occluded. For methods based on global image descriptors, such as the BOW approach, these factors cause the presence of the car to contribute little to the feature vector describing the image. In the case of the car class, the high ranked negative images (Fig. 10c) show an "intuitive" confusion – the first five images shown all contain buses or lorries (not considered part of the "car" class). This may be considered a pleasing result since there is some natural fuzziness in the distinction between the classes "car" and "bus", and the classes certainly share both semantic and visual properties. However, as discussed below, these "natural" confusions are not apparent for all classes.

Fig. 11 shows the five highest ranked positive and negative images for the "cat" class. Here also the confusion appears natural, with all five of the highest ranked non-cat images containing dogs. However, it can also be seen that the *composition* of the images for the cat and dog classes is very similar, and this may play a significant role in the learnt classifiers. This is a bias in the content of flickr images, in that photographers appear to take many "close-up" images of their pets.

Fig. 12 shows corresponding images for the "bicycle" class. The high ranked positive images show a similar pattern to the "car" class, containing uncluttered images of bicycles in "canonical" poses, and images where the scene is dominated by multiple bicycles. For this class, however, the high ranked negative images (Fig. 12b) are anything but intuitive – all of the first five negative images show scenes of birds sitting on branches, which do not resemble bicycles at all to a human observer. The reason for this confusion might be explained by the lack of informative spatial information in current methods. Examining the negative images (Fig. 12b), which are dominated by "wiry" or bar-like features, it seems clear that a BOW representation may closely resemble that of a bicycle, with the representation of branches matching that of the bicycle tubes and spokes. This is a limitation of BOW methods, in that information about the spatial arrangement of features is discarded, or represented only very weakly and implicitly, by features representing the conjunction of other features. For methods using tiling or spatial pyramids, the spatial information is captured

(a) **car**: highest ranked positive images



(b) **car**: lowest ranked positive images



(c) **car**: highest ranked negative images



Fig. 10 Ranked images for the "car" classification task (see text for details of ranking method). (a) five highest ranked positive images (containing cars); (b) five lowest ranked positive images (containing cars); (c) five highest ranked negative images (not containing cars). The number in each image indicates the corresponding median rank.

(a) **cat**: high ranked positive images

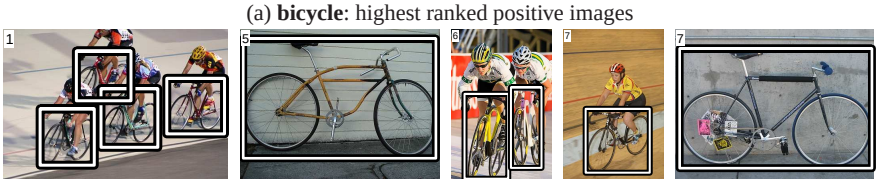


(b) **cat**: high ranked negative images



Fig. 11 Ranked images for the "cat" classification task. (a) five highest ranked positive images (containing cats); (b) five highest ranked negative images (not containing cats). The number in each image indicates the corresponding median rank.

(a) **bicycle**: highest ranked positive images



(b) **bicycle**: highest ranked negative images



Fig. 12 Ranked images for the "bicycle" classification task. (a) five highest ranked positive images (containing bicycles); (b) five highest ranked negative images (not containing bicycles). The number in each image indicates the corresponding median rank.

(a) **chair**: highest ranked positive images



(b) **chair**: highest ranked negative images



Fig. 13 Ranked images for the "chair" classification task. (a) five highest ranked positive images (containing chairs); (b) five highest ranked negative images (not containing chairs). The number in each image indicates the corresponding median rank.

(a) **person**: highest ranked positive images



(b) **person**: highest ranked negative images

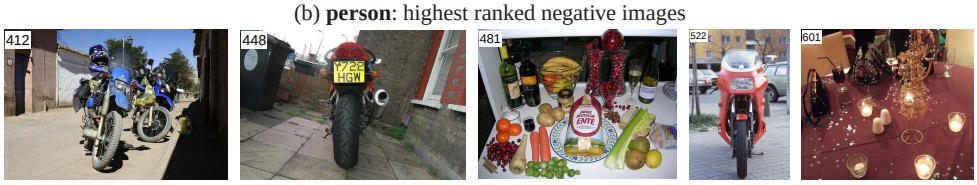


Fig. 14 Ranked images for the "person" classification task. (a) five highest ranked positive images (containing people); (b) five highest ranked negative images (not containing people). The number in each image indicates the corresponding median rank.

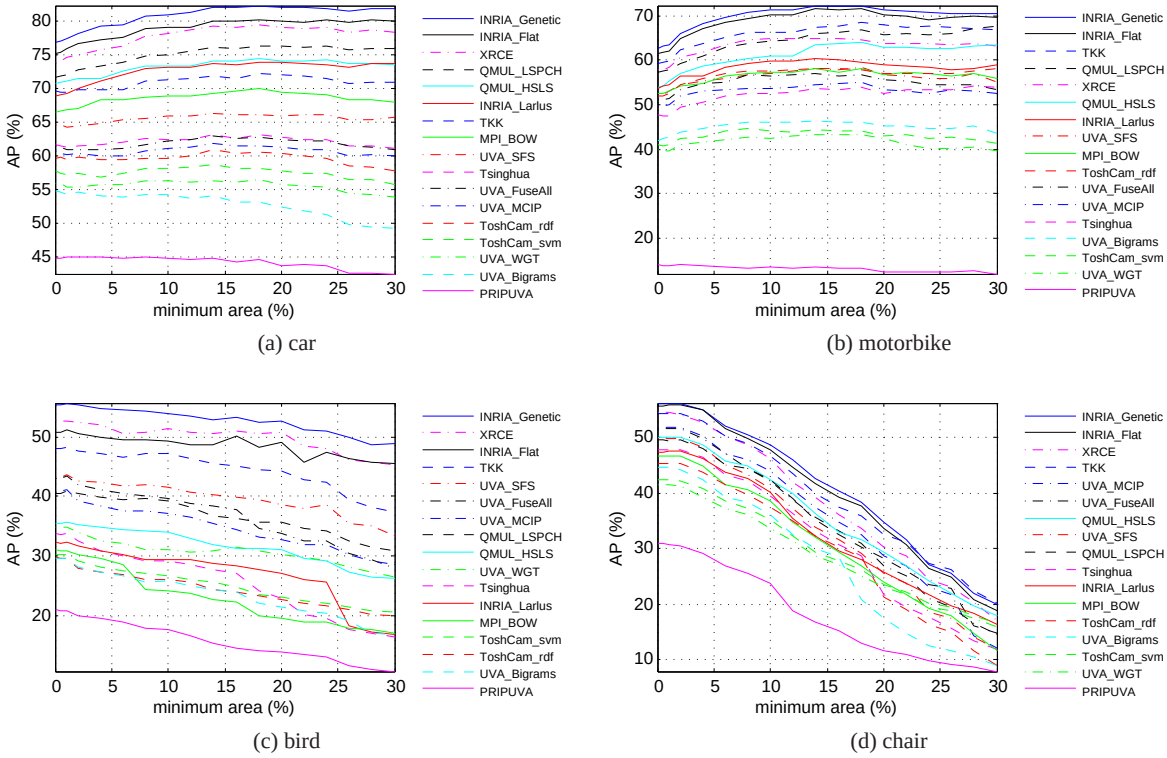
only at a coarse image/scene level, and not at the level of individual objects.

Fig. 13 shows images for the "chair" class. Results here are interesting in that none of the high ranked positive images are close-up views of isolated chairs, even though these are present in the dataset. All the high ranked negative images (Fig. 13b) show indoor scenes which might well be expected to contain chairs, but do not. Only one of the first five negative images contains a sofa, which might be considered the most easily confused class both semantically and visually. It seems likely that in this case the classifiers are learning about the scene *context* of chairs rather than modelling the appearance of a chair itself. Again, this is a somewhat natural consequence of a global classification approach. The use of context may be seen in both positive and negative lights – while there is much interest in the field in exploiting contextual cues to object recognition (Torralba 2003; Sudderth et al 2008; Hoiem et al 2006), the incorporation of context by use of a global descriptor leads to failure when objects are presented out of context, or over-reliance on context when the training set contains mostly images of scenes rather than individual objects. The question of whether current methods are learning object or scene representations is considered further below.

Finally, Fig. 14 shows images for the "person" class. In this case, the negative images (Fig. 14b) contain (i) dining tables (3rd and 5th image), and (ii) motorbikes (1st, 2nd and 4th images). The confusion with the "dining table" class seems natural, in the same manner as the "chair" class, in that the presence of a dining table seems a good predictor for the presence of a person. Statistics of the dataset reveal that the presence of a motorbike is a similarly effective predictor: 68.9% of the images containing motorbikes also contain people (although an alternative explanation may be the elongated vertical shape of the motorbikes seen from the front or rear). These "unintentional" regularities in the dataset are a limiting factor in judging the effectiveness of the classification methods in terms of *object* recognition rather than image retrieval. The *detection* task, see Sect. 6.2, is a much more challenging test of object recognition.

6.1.5 Effect of object size on classification accuracy

All of the methods submitted are essentially *global*, extracting descriptors of the entire image content. The ranking of



(a) car (b) motorbike

(c) bird (d) chair

Fig. 15 Classification results as a function of object size. Plots show results for four representative classes. For each plot the *x*-axis shows the lower threshold on object size for a positive image to be included in the test set; the *y*-axis shows the corresponding AP. A threshold of 30% means that all positive images (e.g. containing "car") which contained fewer than 30% positive (i.e. car) pixels were removed from the test set; a threshold of 0% means that no images were discarded.

images also suggests that the image context is used extensively by the classifiers. It is interesting therefore to examine whether methods are biased toward images where the object of interest is large, or whether conversely the presence of adequate scene context determines the accuracy of the results.

We conducted experiments to investigate the effect of object size on the submitted methods' accuracy. A series of test sets was made in which all positively-labelled images contained at least some proportion of pixels labelled as the object of interest. For example, given a threshold of 10%, only images for which at least 10% of the pixels were "car" were labelled as containing a car; images with some car pixels, but less than 10%, were removed from the test set, and the negative examples always had zero car pixels. The proportion of pixels belonging to a particular class was approximated by the union of the corresponding bounding boxes. Fig. 15 shows results of the experiments for a representative set of classes. For each plot, the *x*-axis shows the threshold on the proportion of positive pixels in an image for the image to be labelled positive, defining one of the series of test sets. For each threshold, the AP was measured and is shown on the *y*-axis.

Fig. 15a/b show classes "car" and "motorbike", for which some positive correlation between object size and AP can be observed. As shown, the AP increases as images with fewer object pixels are discarded, and peaks at the point where the only positive images included have at least 15% object pixels. This effect can be accounted for by the use of interest point mechanisms to extract features, which break down if the object is small such that no interest points are detected on it, and in the case of dense feature extraction, by the dominance of the background or clutter in the global representation, which "swamps" the object descriptors. For the "motorbike" class, the AP is seen to decrease slightly when only images containing at least 20% of object pixels are included – this may be explained by the reduction of relevant context in such images.

For most classes, zero or negative correlation between object size and AP was observed, for example "bird", shown in Fig. 15c. This is compatible with the conclusions from examining ranked images, that current methods are making substantial use of image composition or context. For some classes, e.g. "chair", shown in Fig. 15d, this effect is quite dramatic – for this class the learnt classifiers are very poor at recognising images where chairs are the dominant part of the scene. These results are in agreement with the ranked images shown in Fig. 13b, suggesting a reliance on scene context.

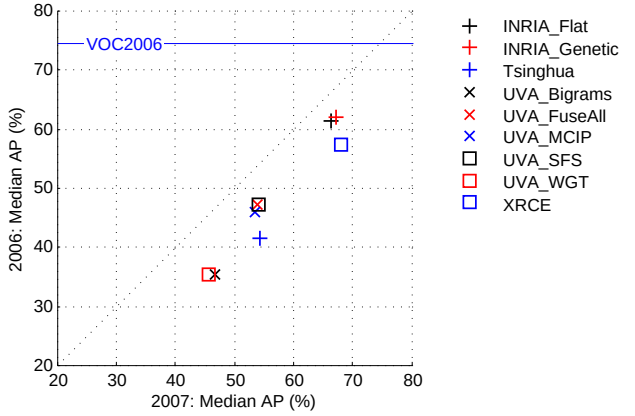


Fig. 16 Comparison of classification results on VOC2006 vs. VOC2007 test sets. All methods were *trained* on the VOC2007 training/validataion set. The median AP over the 10 classes common to the VOC2006/VOC2007 challenges is plotted for each method and test set. The line marked VOC2006 indicates the best result obtained in the VOC2006 challenge, trained using the *VOC2006* training/validation set.

6.1.6 Classification results on the VOC2006 test set

In order to gauge progress since the 2006 VOC challenge, participants in the 2007 challenge were asked to additionally submit results on the *VOC2006* dataset. Since the training phase for many methods is computationally expensive, for example requiring multiple cross-validation runs, participants were asked to submit results trained using *VOC2007* data i.e. not requiring re-training. This allows us to answer two questions: (i) do methods which perform well on the VOC2007 test set also perform well on VOC2006 test set? (ii) do newer methods trained on VOC2007 data outperform older methods trained on VOC2006 data?

Participants submitted classification results on both test sets for 9 methods. Fig. 16 summarises the results. The *x*-axis show the median AP over all classes on the VOC2007 test set. The *y*-axis shows the median AP over all classes on the VOC2006 test set. The line labelled "VOC2006" indicates the best result reported in the 2006 challenge. Note that since the median is taken over the 10 classes common to the 2006 and 2007 challenges (see Fig. 2), the ranking of methods does not match that shown in Fig. 7, for example the *XRCE* method outperforms the *INRIA* methods on the VOC2007 data for this subset of classes.

As the figure shows, there is very high correlation between the results on the VOC2006 and VOC2007 data. This suggests that methods performing well on VOC2007 are "implicitly" more successful, rather than obtaining good results due to excessive fitting of the statistics of the VOC2007 data. There are small differences in the ranking of methods, for example the *XRCE* method is 1st on VOC2007 (over the subset of 10 classes common to both challenges) but 3rd on VOC2006. The *INRIA_Genetic* method gives results marginally lower than *XRCE* on the VOC2007 data, but convincingly better results on the VOC2006 data.

However, for all methods the performance on the VOC2006 data is less than on VOC2007, by 5.0% (*INRIA_FLAT*) to 12.7% (*Tsinghua*). This implies that methods have failed to generalise to the VOC2006 data to some extent. There are two possible reasons: (i) there is fundamentally insufficient variability in the VOC2007 data to generalise well to the VOC2006 data; (ii) the classifiers have overfit some "peculiarities" of the VOC2007 data which do not apply to the VOC2006 data. Factors might include the difference in the time of year of data collection (see Sect. 3.1). A possible concern might be that the better results obtained on VOC2006 are due to the presence of "near-duplicate" images spanning the training/test sets. This possibility was minimised by removing such images when the dataset was collected (see Sect. 3.1).

Tested on the VOC2006 data, the maximum median-AP achieved by a method submitted in 2007 was 62.1% compared to 74.5% in 2006. This again suggests either that the 2007 methods over-fit some properties of the VOC2007 data, or that there were peculiarities of the VOC2006 data which methods trained on that data in 2006 were able to exploit. One such possibility is the inclusion of the MSR Cambridge images in the VOC2006 data (see Sect. 3.1) which may have provided a "boost" to 2006 methods learning their specific viewpoints and simple scene composition.

6.2 Detection

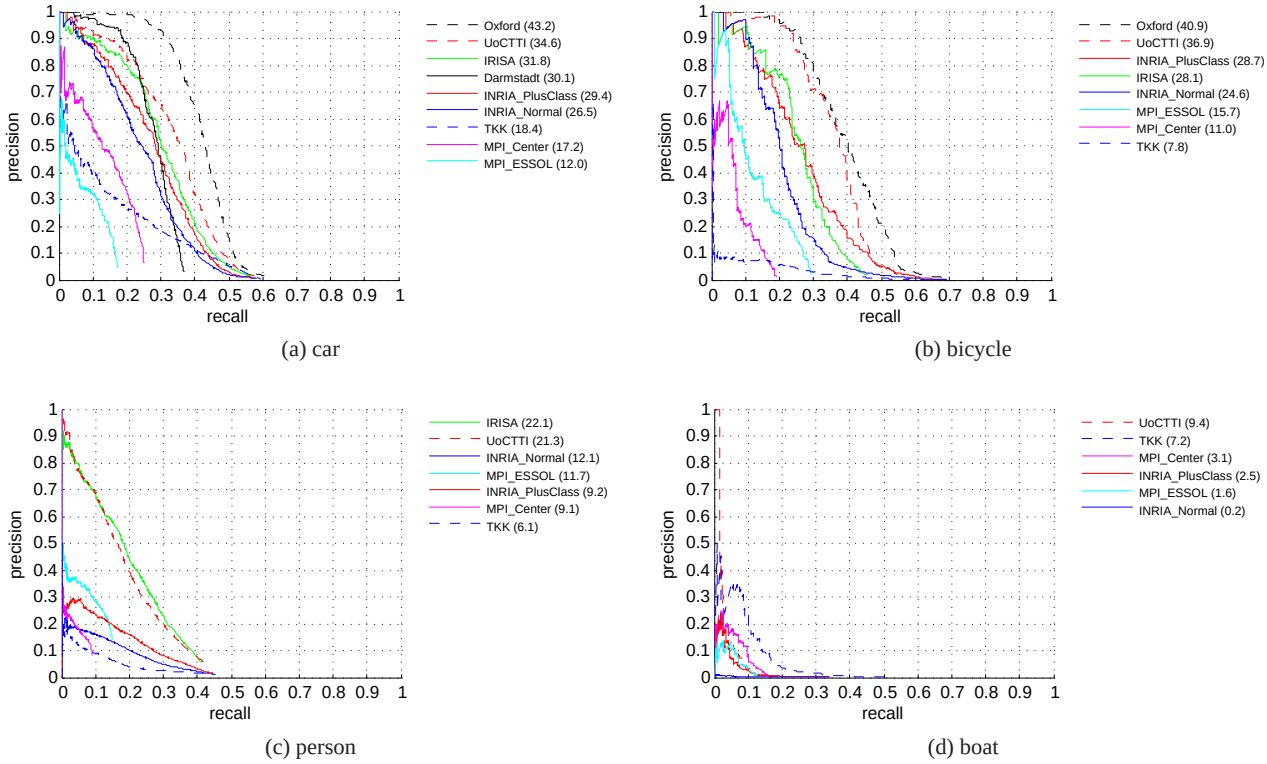
This section reports and discusses the results of the detection task. A total of 9 methods were evaluated. Six participants tackled all of the 20 classes, with the others submitting results on a subset of classes. Table 6 lists the AP for all submitted methods and classes. For each class the method obtaining the greatest AP is identified in bold, and the methods with 2nd and 3rd greatest AP in italics. Precision/recall curves for a representative sample of classes are shown in Fig. 17.

6.2.1 Detection results by method

It is difficult to judge an overall "winner" in the detection task because different participants tackled different subsets of classes (this is allowed under the rules of the challenge). *Oxford* won on all 6 (vehicle) classes that they entered, *UoCTTI* won on 6 classes, and *MPI_ESSOL* on 5. The *Oxford* method achieved the greatest AP for all of the six classes entered, with the AP substantially exceeding the second place result, by a margin of 4.0–11.4%. The *UoCTTI* method entered all 20 classes, and came first or second in

Table 6 Detection results. For each object class and submission, the AP measure (%) is shown. Bold entries in each column denote the maximum AP for the corresponding class. Italic entries denote the results ranked second or third. Note that some participants submitted results for only a subset of the 20 classes.

	aeroplane	bicycle	bird	boat	bottle	bus	car	cat	chair	cow	diningtable	dog	horse	motorbike	person	pottedplant	sheep	sofa	train	tvmonitor
Darmstadt	–	–	–	–	–	–	30.1	–	–	–	–	–	–	–	–	–	–	–	–	–
INRIA_Normal	9.2	24.6	1.2	0.2	*6.8*	19.7	26.5	1.8	*9.7*	3.9	1.7	1.6	*22.5*	15.3	*12.1*	*9.3*	0.2	*10.2*	15.7	24.2
INRIA_PlusClass	13.6	*28.7*	4.1	2.5	*7.7*	*27.9*	29.4	*13.2*	*10.6*	*12.7*	*6.7*	7.1	**33.5**	*24.9*	9.2	*7.2*	1.1	9.2	*24.2*	*27.5*
IRISA	–	28.1	–	–	–	–	*31.8*	2.6	*9.7*	*11.9*	–	–	*28.9*	22.7	**22.1**	–	**17.5**	–	–	*25.3*
MPI_Center	6.0	11.0	2.8	*3.1*	0.0	16.4	17.2	*20.8*	0.2	4.4	4.9	*14.1*	19.8	17.0	9.1	0.4	*9.1*	3.4	*23.7*	5.1
MPI_ESSOL	15.2	15.7	**9.8**	1.6	0.1	18.6	12.0	**24.0**	0.7	6.1	**9.8**	**16.2**	3.4	20.8	11.7	0.2	4.6	**14.7**	11.0	5.4
Oxford	**26.2**	**40.9**	–	–	–	**39.3**	**43.2**	–	–	–	–	–	–	**37.5**	–	–	–	–	**33.4**	–
TKK	*18.6*	7.8	*4.3*	*7.2*	0.2	11.6	18.4	5.0	2.8	10.0	*8.6*	*12.6*	18.6	13.5	6.1	1.9	3.6	5.8	6.7	9.0
UoCTTI	*20.6*	*36.9*	*9.3*	**9.4**	**21.4**	*23.2*	*34.6*	9.8	**12.8**	**14.0**	0.2	2.3	18.2	*27.6*	*21.3*	**12.0**	*14.3*	*12.7*	13.4	**28.9**



(a) car (b) bicycle

(c) person (d) boat

Fig. 17 Detection results. Precision/recall curves are shown for a representative sample of classes. The legend indicates the AP (%) obtained by the corresponding method.

14. *MPI_ESSOL* also entered all 20 classes, but it is noticeable that on some classes the AP score for this method is poor relative to other entries.

These methods differ quite significantly in approach: *Oxford* used interest point detection to select candidate detection regions, and applied an SVM classifier using a spatial pyramid (Lazebnik et al 2006) representation to the candidate region; the *UoCTTI* method used a sliding window approach, but with a "star" model of parts; and *MPI_ESSOL* used an SVM classifier based on a BOW or spatial pyramid representation of the candidate window.

It would have been particularly interesting to see results of the *Oxford* method on all classes, since it might be expected that the use of interest points and a fixed grid representation might limit the applicability to classes with limited distinguishing features or significant variability in shape, e.g. animals.

Promising results were obtained by all methods, but with results for each method varying greatly among the object classes. It seems clear that current methods are more or less suited to particular classes. An example is the failure of the *UoCTTI* method on the "dog" class (AP=2.3) compared to the *MPI_ESSOL* method (AP=16.2). While the former emphasises shape, the latter uses a BOW/spatial pyramid representation which might better capture texture, but captures shape more coarsely. Conversely, the *UoCTTI*

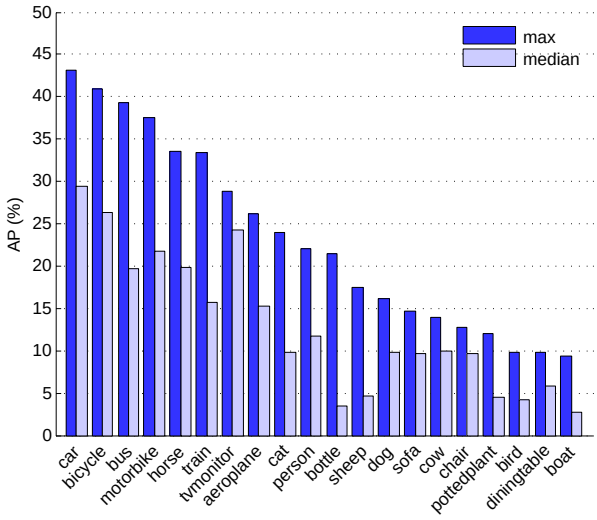


Fig. 18 Summary of detection results by class. For each class two values are shown: the maximum AP obtained by any method (max) and the median AP over all methods (median).

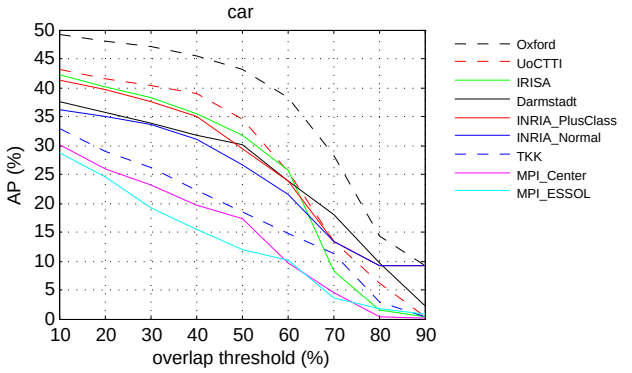


Fig. 19 Effect of the bounding box overlap threshold on the AP measure. For the class "car", on which the submitted methods gave the best results, AP is plotted as a function of the overlap threshold. The threshold adopted for the challenge is 50%.

method obtains good results on "bottle" (AP=21.4), where it is expected that shape is a very important feature, and the *MPI_ESSOL* method fails (AP=0.1). The trained detector used by the *UoCTTI* method (Felzenszwalb et al 2008) has been made publicly available, and would form a reasonable state-of-the-art baseline for future challenges.

6.2.2 Detection results by class

Fig. 18 summarises the results obtained by object class, plotting for each class the maximum and median AP taken over all methods. Results are shown ordered by decreasing maximum AP. It should be noted that, because some participants only submitted results for some classes, the number of results available varies for each class. There is substantial variation in the maximum AP as a function of object class, from 9.4% (boat) to 43.2% (car). The median AP varies from 2.8% (boat) to 29.4% (car). The median results can be seen to approximately follow the ranking of results by maximum AP.

Results on some classes e.g. car/bicycle (Fig. 17a-b) are quite promising, with the best performing methods obtaining precision close to 100% for recall up to 15–20%. However, precision drops rapidly above this level of recall. This suggests that methods are retrieving only a subset of examples with any accuracy, perhaps the "canonical" views (e.g. car side, car rear). A challenge for future methods is to increase the recall. In the related domain of face detection the move from frontal-only detection to arbitrary pose has proved challenging.

It can be seen from Fig. 18 that the best results were obtained for classes which have traditionally been investigated in object detection, e.g. car, bicycle and motorbike. Such classes have quite predictable visual properties, with distinctive parts e.g. wheels, and relatively fixed spatial arrangement of parts. For classes with significant variation in shape or appearance e.g. people (Fig. 17c) and household objects (which are often significantly occluded), results are substantially worse. Results on the important "person" class were, however quite promising overall. The best results in terms of AP on this class were obtained by the *IRISA* and *UoCTTI* methods. As noted in Sect. 5.2 the *IRISA* method trained multiple person detectors, for example "person on horse/person on bicycle". The *UoCTTI* method is also potentially better able to deal with varying articulation, by its approach of simultaneous "pose" inference and detection.

For several classes the results are somewhat counterintuitive, for example good results are obtained on the "train" class (max AP=33.4%) which might be expected to be challenging due to the great variation in appearance and aspect ratio with pose. The results for this class may be explained by the inclusion of several methods which exploited whole image classification – *MPI_Center* which predicts a single detection per image of fixed size, and *INRIA_PlusClass* which combines sliding window detection with a global classifier. Because trains tend to appear large in the image, these global methods prove successful on this data, however it is notable that the *Oxford* method also did well for this class. For the "horse" class, the good results may be attributable to unwanted regularities in the dataset, which includes many images of horses taken by a single photographer at a single gymkhana event. Such regularities will be reduced in the VOC2008 dataset by distributing searches over time, as noted in Sect. 3.1. Results for the classes with low AP, for example boat (Fig. 17d) leave much scope for improvement, with precision dropping to zero by 20% recall. The VOC2007 dataset remains extremely challenging for current detection methods.

(a) **car** true positives: *Oxford* method



(b) **car** true positives: *UoCTTI* method



(c) **car** true positives: *IRISA* method

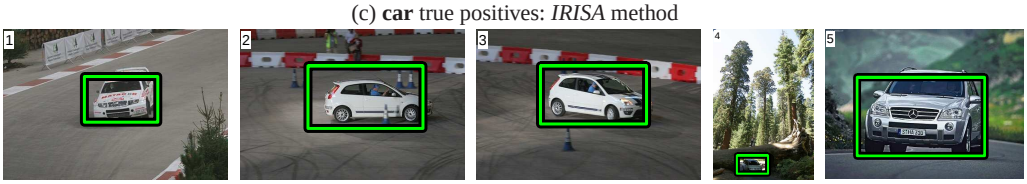
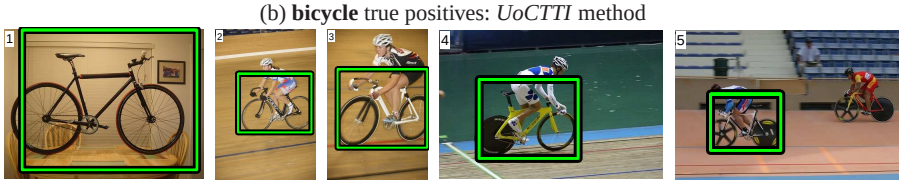


Fig. 20 Highest ranked true positive detections for the "car" detection task. The five highest ranked true positives are shown for each of the three methods with greatest AP. The number in each image indicates the rank of the detection.

(a) **bicycle** true positives: *Oxford* method



(b) **bicycle** true positives: *UoCTTI* method



(c) **bicycle** true positives: *INRIA_PlusClass* method



Fig. 21 Highest ranked true positive detections for the "bicycle" detection task. The five highest ranked true positives are shown for each of the three methods with greatest AP. The number in each image indicates the rank of the detection.

6.2.3 Evaluation of the overlap threshold

As noted in Sect. 4.2, detections are considered true positives if the predicted and ground truth bounding boxes overlap by 50% according to the measure defined in Eqn. 3, with the threshold of 50% set low to account for uncertainty in the bounding box annotation. We evaluated the effect the overlap threshold has on the measured AP by varying the threshold. Fig. 19 shows AP as a function of the overlap threshold for the class "car", on which the best results (in terms of AP with overlap threshold of 50%) were obtained by the submitted methods. One caveat applies: participants were aware of the 50% threshold, and were therefore free to optimise their methods at this threshold, for example in their schemes for elimination of multiple detections.

As Fig. 19 shows, the measured AP drops steeply for thresholds above 50%, indicating that none of the methods give highly accurate bounding box predictions. Reducing the threshold to 10% results in an increase in measured AP of around 7.5%. Note that for all but one pair of methods (*Darmstadt* and *INRIA_PlusClass*) the ordering of methods by AP does not change for any threshold in the range 0–50%

(a) **bicycle** false positives: *Oxford* method



(b) **bicycle** false positives: *UoCTTI* method



(c) **bicycle** false positives: *INRIA_PlusClass* method



Fig. 22 High ranked false positive detections for the "bicycle" detection task. The false positives shown are in images where *no* bicycles are present. The number in each image indicates the rank of the detection. Results are shown for the three methods with greatest AP.

(a) **motorbike** false positives: *Oxford* method



(b) **motorbike** false positives: *UoCTTI* method



(c) **motorbike** false positives: *INRIA_PlusClass* method



Fig. 23 High ranked false positive detections for the "motorbike" detection task. The false positives shown are in images where *no* motorbikes are present. The number in each image indicates the rank of the detection. Results are shown for the three methods with greatest AP.

(the AP of these two methods at a threshold of 50% differs by only 0.7%). We conclude that the measure is performing satisfactorily, capturing the proportion of objects detected without overly penalising imprecise bounding box predictions.

6.2.4 What are the detection methods learning?

As in the classification task it is interesting to examine the success and failure modes of the methods to derive some insight into what current methods are learning, and what limitations might be addressed in the development of future methods.

Each detection method provides a list of bounding boxes ranked by the corresponding confidence output. We present some of the highest ranked true positive (object) and false positive (non-object) detections here. Since the detection methods varied greatly in approach and success, as measured by AP, we present individual results for selected classes and methods. For a given class, we present results of the three methods giving greatest AP. The classes selected were those where results are particularly promising, or in-

(a) **person** false positives: *IRISA* method



(b) **person** false positives: *UoCTTI* method



(c) **person** false positives: *INRIA_Normal* method



Fig. 24 High ranked false positive detections for the "person" detection task. The false positives shown are in images where *no* people are present. The number in each image indicates the rank of the detection. Results are shown for the three methods with greatest AP.

(a) **person** "near misses": *IRISA* method



(b) **person** "near misses": *UoCTTI* method



(c) **person** "near misses": *INRIA_Normal* method



Fig. 25 "Near miss" false positives for the "person" detection task. The images shown contain people, but the detections do not satisfy the VOC overlap criterion. The number in each image indicates the rank of the detection. Results are shown for the three methods with greatest AP.

teresting observations can be made e.g. confusion between "motorbike" and "bicycle".

Fig. 20 shows the five highest ranked true positive detections for the "car" class. The methods shown, obtaining greatest AP, are *Oxford*, *UoCTTI* and *IRISA*. As can be seen the *Oxford* method has a preference for large objects (Fig. 20a) which is less apparent for the other two methods (Fig. 20b/c). We analyse this bias toward large objects further in the next subsection. For this class, there is no apparent preference for a particular viewpoint or "aspect" – the methods all return cars with varying pose at high confidence. However, for the "bicycle" class shown in Fig. 21, the most successful methods (*Oxford*, *UoCTTI* and *INRIA_PlusClass*) all return side views of bicycles with highest confidence. For the *UoCTTI* method in particular there seems no preference for left or right facing bicycles, though the bias toward right facing bicycles for the *Oxford* method may be a statistical bias in the dataset.

We turn now to the false positive detections – bounding boxes which do not correspond to an object of interest. For each class and method we show high-ranked false positive detections. To increase the diversity of results presented

we have filtered the images shown: (i) images with *any* (undetected) object of interest have been removed, though see discussion of the "person" class below; (ii) only the most confident false positive detection per image is shown. For example in Fig. 22b, multiple high confidence false positives were output for the 3rd image due to the repeated structure.

Fig. 22 shows false positives for the "bicycle" class output by the *Oxford*, *UoCTTI* and *INRIA_PlusClass* methods. All methods generated some "intuitive" false positives on motorbikes, though in many such cases the predicted bounding box does not correspond to the full extent of the object. It is interesting to observe that several of the false positives for the *Oxford* method are the same images of tree branches which confused the classification methods (Fig. 12b). This may be explained by the pyramid representation of spatial information in this method or by the method learning the strong gradients of the frames (if you squint you can see a frame in the second and fourth images of Fig. 22a). For the *UoCTTI* method, the false positives were observed to often resemble the side-view of a bicycle as "two blobs horizontally aligned" (Fig. 22b, 3rd and 4th image). The most confident false positive output by this method is actually a drawing of a bicycle on a traffic sign, not labelled as "bicycle" according to the annotation guidelines. For the *INRIA_PlusClass* method, 4 of the 5 highest confidence false positive images contain motorcycles, however the poor prediction of bounding boxes in these cases suggest that the scene context introduced by the incorporation of a global image classifier in this method may be a factor, rather than a "natural" confusion between the classes.

Fig. 23 shows corresponding high ranked false positives for the "motorbike" class, with the same methods as for the "bicycle" class shown. For the *UoCTTI* method, 4 out of 5 false positives are bicycles, with the remaining false positive shown covering a pair of car wheels. These results suggest that this method is really capturing something about the dominant shape of the motorbike. The *Oxford* method outputs two bicycles in the first five false positives, and the *INRIA_PlusClass* method outputs one. The remaining high ranked false positives for these two methods are difficult to explain, consisting mainly of highly cluttered scenes with little discernable structure.

Fig. 24 shows high ranked false positives for the "person" class, with the three most successful methods shown: *IRISA*, *UoCTTI* and *INRIA_Normal*. This class is particularly challenging because of the high variability in human pose exhibited in the VOC dataset. As can be seen, it is difficulty to see any consistent property of the false positives. Some bias toward "elongated vertical" structures can be observed e.g. trees (Fig. 24a) and dogs in a frontal pose (Fig. 24b), and more of these were visible in lower ranked false positives not shown. However, many false positives seem to be merely cluttered windows with strong texture. The fifth false positive output by the *IRISA* method (Fig. 24a) is interesting (motorbike panniers occluded by another motorbike) and is most likely an artefact of that method learning separate "person on X" detectors.

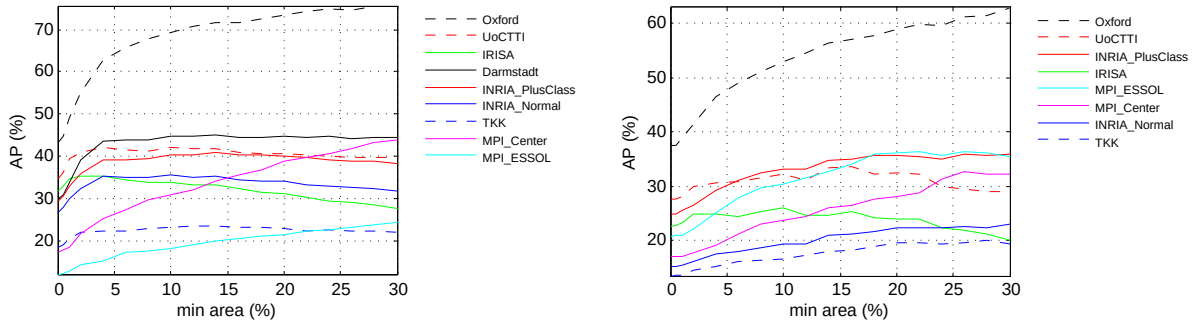
Thus far the false positives shown exclude images where *any* object of interest is present. Fig. 25 shows false positives for the "person" class, where people *are* present in the image. These represent "near miss" detections where the predicted bounding box does not meet the VOC overlap criterion of 50%. As noted, this class presents particular challenges because of the high variability in pose and articulation. As can be seen in Fig. 25a/b, the localisation of these false positives for the *IRISA* and *UoCTTI* methods is generally quite good, e.g. the top of the bounding box matches the top of the head, and the person is horizontally centred in the bounding box. The failure modes here are mainly inaccurate prediction of the vertical extent of the person (Fig. 25a and Fig. 25b, 1st and 2nd image) due e.g. to occlusion, and failure on non-frontal poses (Fig. 25b, 1st and 5th images). This is a limitation of current methods using fixed aspect ratio bounding boxes, which is a poor model of the possible imaged appearance of a person. The false positive in Fig. 25a, 1st image, is accounted for by the "person on X" approach taken by the *IRISA* method. The high ranked near misses output by the *INRIA_Normal* method (Fig. 25c) mostly cover multiple people, and might be accounted for by capturing person "texture" but not layout.

As noted in Sect. 4.2.2, in the 2007 challenge we introduced a "person layout" taster to further evaluate the ability of person detectors to correctly "parse" images of people, and motivate research into methods giving more detailed interpretation of scenes containing people.

6.2.5 Effect of object size on detection accuracy

As in the classification task, it is interesting to investigate how object size affects the accuracy of the submitted detection methods, particularly for those such as the *Oxford* method which makes use of interest point detection, which may fail for small objects, and the *INRIA_PlusClass* method which combines sliding window detection with a whole image classifier.

We followed a corresponding procedure to that for the classification task, creating a series of test sets in which all objects smaller than a threshold area were removed from consideration. For the detection task, this was done by adding a "difficult" annotation to such objects, so that they count neither as false positives or negatives. Fig. 26 shows results of the experiments for a representative set of classes. For each plot, the *x*-axis shows the threshold on object size, as a proportion of image size, for an object to be included in the evaluation. For each threshold, the AP was measured and is shown on the *y*-axis.



(a) car (b) motorbike

Fig. 26 Detection results as a function of object size. Plots show results for two representative classes. For each plot the *x*-axis shows the lower threshold on object size for an object to be included in the test set; the *y*-axis shows the corresponding AP. A threshold of 30% means that all objects smaller than 30% of the image area were ignored in the evaluation (contributing neither true nor false positives); a threshold of 0% means that no objects were ignored.

As Fig. 26 shows, all submitted methods have lower AP for very small objects below around 2% image area. For the "car" class (Fig. 26a), the accuracy of most methods does not increase if objects less than 5% area are removed from the evaluation, indicating limited preference for large objects. For the *INRIA_Normal* and *IRISA* methods the AP can be seen to fall slightly with increasing object size, indicating that some highly ranked correct detections are for small objects. For the "motorbike" class (Fig. 26b), the *INRIA_PlusClass* and *MPI_ESSOL* methods peak for objects above around 17% image area, the *IRISA* and *UoCTTI* methods show no clear preference for large objects, and AP for all other methods increases monotonically with object size.

Three methods show substantial correlation between object size and AP: *MPI_Center*, *MPI_ESSOL* and *Oxford*. The *MPI_Center* method outputs a fixed bounding box with area 51% of the image, and confidence determined by a global image classifier. This clearly biases results to images where most of the image is covered by the object of interest, and while an interesting baseline (as intended), is not a successful strategy since many of the objects in the dataset are small. The *MPI_ESSOL* method has two aspects which may bias it toward larger objects: (i) it combines a whole image classifier with a sliding window detector to "score" detections; (ii) it incorporates a log Gaussian prior on object size, fit by maximum likelihood to the training data, and this prior may have biased results toward large objects. The *Oxford* method relies on scale-invariant interest point operators to provide candidate detections, and the lack of interest points on small objects may explain the correlation between its accuracy and object size.

6.2.6 Detection results on the VOC2006 test set

As in the classification task, participants in the detection task of the 2007 challenge were asked to additionally submit results on the VOC2006 dataset, trained using the VOC2007 data.

Participants submitted classification results on both test sets for 3 methods. Table 7 summarises the results, showing the AP (%) obtained for each method, class and test set. The final row shows the maximum AP obtained by any method in the 2006 challenge, trained on the *VOC2006* data. Since participants submitted results for different subsets of classes, the results have not been summarised e.g. by median AP as in the classification task.

For all but one class the ranking of methods by AP is the same for the VOC2007 and VOC2006 datasets, suggesting that methods performing well on VOC2007 are "intrinsically" more successful, rather than obtaining good results due to excessive fitting of the statistics of the VOC2007 data. For the "cat" class, the *UoCTTI* method comes first and the *Oxford* method second, reversing the order on VOC2007, but the difference in AP is small (53.5% vs. 55.5%).

Particularly encouraging is the observation that for 7 out of 10 classes a method submitted in 2007 achieves greater AP than any of the 2006 methods. The *UoCTTI* method exceeds the 2006 results on 7 of the 10 classes entered, the *Oxford* method on all four classes entered, and the *IRISA* method on 4 of the 8 classes entered. The improvement is substantial, e.g. 19.1% AP on the "bus" class (*Oxford*) and 9.8% on the "person" class (*UoCTTI*). While it is possible that the improvement is due to the VOC2007 training/validation data being more "useful", this effect was not observed for the classification task. It therefore seems likely that the results represent measurable progress in object detection.

6.3 Segmentation

All participants in the detection challenge were automatically entered into the segmentation challenge by deriving

Table 7 Comparison of detection results on VOC2006 vs. VOC2007 test sets. All methods were *trained* on the VOC2007 training/validataion set. The AP measure (%) is shown for each method, class, and test set. The final row (VOC2006) lists the best result obtained in the VOC2006 challenge, trained using the *VOC2006* training/validation set. Bold entries denote the maximum AP for each dataset and class. Bold entries in the final row indicate where results obtained in 2006 exceeded those obtained in 2007.

		bicycle	bus	car	cat	cow	dog	horse	motorbike	person	sheep
Test on 2007	**IRISA**	28.1	–	31.8	2.6	11.9	–	**28.9**	22.7	**22.1**	**17.5**
	Oxford	**40.9**	**39.3**	**43.2**	–	–	–	–	**37.5**	–	–
	UoCTTI	36.9	23.2	34.6	**9.8**	**14.0**	**2.3**	18.2	27.6	21.3	14.3
Test on 2006	**IRISA**	35.2	–	48.2	9.4	20.9	–	**18.3**	33.3	21.1	**26.2**
	Oxford	**56.8**	**36.0**	53.5	–	–	–	–	**53.9**	–	–
	UoCTTI	56.2	23.6	**55.5**	**10.3**	**21.2**	**9.9**	17.3	43.9	**26.2**	22.1
VOC2006	**Best**	44.0	16.9	44.4	**16.0**	**25.2**	**11.8**	14.0	39.0	16.4	25.1

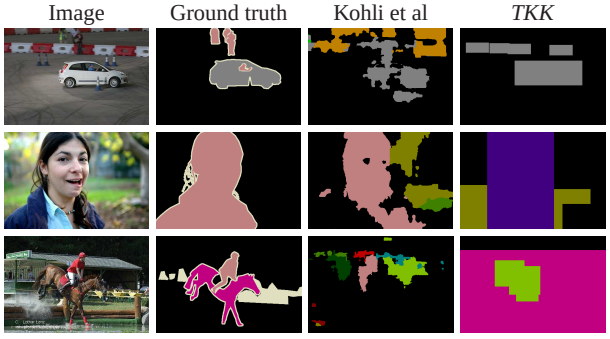


Fig. 27 Example segmentation results. Columns show: test images, ground truth annotations, segmentations from Kohli et al (2008) and segmentations derived from the bounding boxes of the *TKK* detection method.

a segmentation from the inferred bounding boxes (overlaps were resolved heuristically). In addition, only one segmentation-specific method was submitted, by Lubor Ladicky, Pushmeet Kohli and Philip Torr of Oxford Brookes University (Kohli et al 2008). Example segmentations from this team and from the *TKK* automatic entry are shown in Fig. 27. The best overall performance was given by one of the "automatic" participants (segmentation derived algorithmically from detection), most likely due to an unfinished segmentation-only entry. In future challenges, it is anticipated that methods which are optimised for the segmentation problem will outperform automatic detection entries. In any case, providing a challenge which directly compares detection and segmentation methods should help encourage innovation in how to combine these two types of methods to best effect.

6.4 Person Layout

Only one result was submitted for the person layout taster, by Martin Bergtholdt, Jörg Hendrik Kappes and Christoph Schnörr of the University of Mannheim (Bergtholdt et al 2006). Fig. 28 shows some example results for this method.

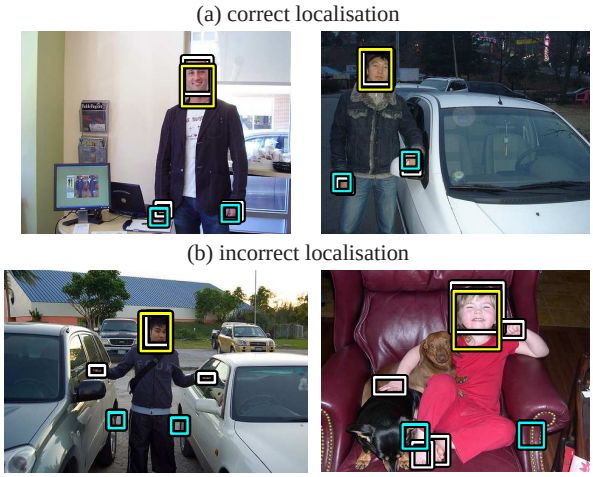


Fig. 28 Example person layout results for the *Mannheim* method. For each image the ground truth bounding boxes are shown in white, and the predicted bounding boxes colour-coded: yellow for "head" and cyan for "hand".

For some images, where the person is in a "canonical" frontal pose, the method successfully localises the parts (Fig. 28a). For more varied poses, the method fails to predict the correct locations, or confuses hands and feet (Fig. 28b). Despite some correct results, the ranking of results by the method's confidence output is poor, such that the measured AP is zero. This raised the question of whether the evaluation criterion adopted for VOC2007 is sufficiently sensitive, and as described in Sect. 4.2.2, the requirements for the person layout taster have been relaxed for the VOC2008 challenge.

7 Discussion and the Future

The VOC challenge has already had a significant, and we believe positive, impact in terms of providing a rich, standardised dataset for the community and an evaluation framework for comparing different methods. Participation in the challenges has increased steadily since their first introduc-

tion, as has the use of the VOC dataset beyond the challenge itself. For example, in CVPR07 the VOC dataset was referenced in 15 papers; in CVPR08 this number increased to 27, i.e. almost a year-to-year doubling of the dataset's popularity. To retain this popularity, the challenge must evolve to meet the requirements and address the criticisms of the research community. In the following sections we discuss some criticisms of the challenge, look at how the challenge is evolving through the taster competitions, and suggest directions in which the dataset and challenge can be improved and extended in the future.

7.1 Criticisms

No benchmark remains without criticism for long, and the VOC challenge has not been an exception. A common objection raised about any competition of this sort is that: "Datasets stifle innovation, because the community concentrates effort on this data to the exclusion of others". While it is difficult to avoid this effect completely, if the challenge is well ahead of capabilities then it will not necessarily stifle the types of methods used. Datasets have a shelf life, and as performance starts to saturate a new one is needed to drive research. Conversely, it is also necessary for datasets to remain consistent, so that they can be used to gauge progress made by the community. Assessing progress is difficult if the test (and training) set are different every time the challenge is run. The VOC challenge aims to meet these apparently contradictory goals of innovation and consistency by introducing separate "taster" challenges to promote research in new directions (see the next section), while retaining the existing classification and detection competitions so that progress can be consistently tracked.

Fostering innovation is also a question of the attitude of the community as a whole: it is important that we do not discourage novel approaches to object recognition simply because they do not yet achieve the greatest success as measured by our benchmarks. High methodological novelty must not be sacrificed on the altar of benchmark ranking, and this is the last thing the VOC challenge is intended to achieve. An important part of encouraging novel methods is our selection of speakers for the annual challenge workshop, where we have given time to both particularly successful, and particularly interesting methods.

A further criticism raised against the VOC series of challenges in particular is that the level of difficulty is too high, thereby obscuring the way forward. However, methods submitted in 2007 for the detection task demonstrated substantial improvements over those submitted in 2006 (see Sect. 6.2.6). We are of the opinion that providing researchers with such challenging, yet natural, data is only stimulating progress. It is the very fact of being well ahead of current capabilities which makes the dataset so useful. In contrast, datasets for which performance is "saturated" are likely to encourage fine tuning of implementation details rather than fundamental progress, and such progress may be unmeasurable, being lost in the noise.

A fundamental question is whether the VOC challenges are probing for the right kind of tasks. In a recent paper, Pinto et al (2008) criticised the use of "natural" images altogether, arguing for the use of synthetic data (e.g. rendered 3D models) for which one has better control over the variability in the data – parameter settings can be sampled at will and annotation is not needed, as perfect ground truth is available by design. In their view, this is a much better way to generate the variability that is needed to critically test recognition performance. This issue of whether to use natural images or completely control imaging conditions is an ongoing debate in the psychophysics community. In our case, the VOC datasets have been designed to contain large variability in pose, illumination, occlusion, etc. Moreover, correlations that occur in the real world are captured, whereas synthetic datasets cannot be expected to reflect those faithfully.

7.2 Taster competitions

The taster competitions, which make demands of methods quite far ahead of the state-of-the-art, aim to play a key part in encouraging fundamental research progress. These were introduced in 2007 to encourage both diversity of approach and the development of more powerful methods to address these more demanding tasks. For example, the segmentation competition not only requires much more precise localisation of objects than the detection task, but it has also been set up to allow either detection-based or segmentation-based approaches to be used. The hope is that the two approaches are complementary, so that detection methods can be used to improve segmentation performance and vice-versa. This belief is justified by the similar situation which has already arisen between the classification and detection tasks, where global image classification has aided detection performance (see Sect. 5.2). By encouraging participants to blend the best aspects of different methodologies, a greater diversity of approaches will be encouraged.

It is inevitable that any challenge is very much of its time, only testing what can be thought of by current practitioners, governed by current methods and hardware, and to some extent unaware of these limitations. Through the use of the taster competitions, the VOC challenge is being updated to allow a broader range of approaches and to address more current research issues. However, it is recognised that the challenge must continue to adapt and remain agile in responding to the needs and concerns of the growing community of researchers who use the datasets and participate in the competitions.

7.3 The Future

In the area of object class recognition, a lot of progress is being made and the requirements for a benchmark evolve quickly with this evolution. Here we give a non-exhaustive list of aspects which could be improved or added in future VOC challenges.

More object classes. A first and rather obvious extension is to increase the number of annotated object classes. A primary goal here is to put more emphasis on the issue of scalability – running as many detectors as there are object classes may not remain a viable strategy, although this is by far the dominant approach today. Different aspects of detection schemes may become important, for example the ability to share features between classes (Torralba et al 2007), or exploit properties of multiple "parent" classes (Zehnder et al 2008). Introducing more classes would also stimulate research in discrimination between more visually similar classes, and in exploiting semantic relations between classes, for example in the form of a class hierarchy. However, increasing the number of classes will also pose additional difficulties to the running of the VOC challenge: (i) it will prove more difficult to collect sufficient data per class; (ii) it raises questions of how to annotate objects accurately, for example labelling an object as "van" vs. "truck" is often subjective; (iii) evaluation of recognition must be more flexible, for example a method might assign a class from $\{hatchback, car, vehicle\}$ and be assigned varying "scores" dependent on accuracy or level of detail.

Object parts. VOC2007 introduced annotation of body parts in order to evaluate and encourage development of methods capable of more detailed image annotation than object location alone. Such more detailed indication of the parts of objects is an important direction to pursue. Although many techniques today start from local features, these features typically have very little to do with the semantic parts of the objects. However, often the purpose of object detection and recognition is to support interaction with objects (e.g. in robotics). A good understanding of where parts are (arms, wheels, keyboards, etc.) is often essential to make such practical use of object recognition, and should be incorporated into at least a component of the evaluation scheme.

Thus far, VOC has confined itself to object classes and annotation where "discrete" objects can be identified. With the introduction of the segmentation taster, it is natural to also include "stuff" classes (grass, sky, etc.) and additionally consider annotation of classes which can appear as "stuff" in the distance e.g. "person" vs. "crowd" – images containing such ambiguities are currently omitted from the VOC dataset.

Beyond nouns. Increasingly, vision researchers are forging strong links with text analysis, and are exploiting tools coming from that area such as WordNet (Fellbaum 1998). Part of this endeavour is to build vision systems that can exploit and/or generate textual descriptions of scenes. This entails bringing objects (nouns) in connection with actions (verbs) and attributes (adjectives and adverbs). As progress in this direction continues, it will be appropriate to introduce benchmarks for methods producing richer textual descriptions of a scene than the "noun + position" outputs which are currently typical. The interest in methods for exploiting textual description at training time also suggests alternative *weaker* forms of annotation for the dataset than bounding boxes; we discuss this further below.

Scene dynamics. Thus far, the VOC challenge has focused entirely on classifying and detecting objects in still images (also the case for VOC2008). Including video clips would expand the challenge in several ways: (i) as training data it would support learning richer object models, for example 3D or "multi-aspect". Video of objects with varying viewing direction would provide relations between parts implicitly available through tracking; (ii) as test data it would enable evaluation of new tasks: object recognition from video (e.g. people), and recognition of actions. This would also bring the VOC challenge into the domain of other benchmarks, e.g. TRECVID which includes an "interactive search" task with increasing emphasis on events/actions such as "a door being opened" or "a train in motion".

Alternative annotation methods. Manual annotation is time-consuming and therefore expensive. For example, annotation of the VOC2008 dataset required around 700 person hours. Moreover, since the VOC challenge runs annually, new test data is required each year in order to avoid participants having access to the ground truth annotations and over-fitting on the test set. Increasing the level of annotation, for example by increasing the number of classes, only makes annotation more time-consuming.

We also found that when increasing the number of classes, from 10 in 2006 to 20 in 2007, annotators made many more mistakes as they failed to hold in memory the complete set of classes to be annotated. This in turn required more time to be allocated to checking and correction to ensure high quality annotation. This raises several questions concerning: how the annotation format relates to ease-of-annotation, how much agreement there is between different human annotators e.g. on bounding box position, and how the annotation tools affect annotation quality. To date, we have not yet gathered data during the checking process that could help answer these questions and this is something we aim to rectify in future years. Annotating pixel-wise segmentations instead of bounding boxes puts even higher pressure on the sustainability of manual annotation. If object

parts, attributes and especially video are to be added in future, then the method of annotation will certainly need to evolve in concert with the annotation itself. Possibilities include recruiting help from a much larger pool of volunteers (in the footsteps of LabelMe), combined with a centralised effort to check quality and make corrections. We are also investigating the use of systems like Mechanical Turk to recruit and pay for annotation (Sorokin and Forsyth 2008; Spain and Perona 2008). Alternatively, commercial annotation initiatives could be considered, like the aforementioned Lotus Hill dataset (Yao et al 2007), in combination with sampled quality inspection.

As noted above, there has recently been considerable interest in learning recognition from "weak" supervision (Duygulu et al 2002; Fergus et al 2007). This suggests alternative forms of annotation which could be introduced, for example per-image annotation with keywords or phrases (e.g. "red car in a street scene"). Such annotation could be provided alone for some images, in addition to a set of images with more precise annotation, providing complementary supervision for training at low cost. Another possibility for "lighter" annotation is to collect (possibly additional) training images directly from a web search engine (such as Google image search) (Fergus et al 2005). The additional complication here is that such data is typically noisy, in that only a subset of the images are relevant to the supplied search terms.

The future is bright. There has been tremendous progress in object class recognition this decade. At the turn of the millennium, few would have dreamt that by now the community would have such impressive performance on both classification and detection for such varied object classes as bicycles, cars, cows, sheep, and even for the archetypal functional class of chairs. This progress has gone hand in hand with the development of image databases — which have provided both the training data necessary for researchers to work in this area; and the testing data necessary to track the improvements in performance. The VOC challenge has played a vital part in this endeavour, and we hope that it will continue to do so.

Acknowledgements

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

We gratefully acknowledge the VOC2007 annotators: Moray Allan, Patrick Buehler, Terry Herbert, Anitha Kannan, Julia Lasserre, Alain Lehmann, Mukta Prasad, Till Quack, John Quinn, Florian Schroff. We are also grateful to James Philbin, Ondra Chum, and Felix Agakov for additional assistance.

Finally we would like to thank the anonymous reviewers for their detailed and insightful comments.

References

Bergtholdt M, Kappes J, Schnörr C (2006) Learning of graphical models and efficient inference for object class recognition. In: Proceedings of the Annual Symposium of the German Association for Pattern Recognition (DAGM06), pp 273–283

Chum O, Zisserman A (2007) An exemplar model for learning object classes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition

Chum O, Philbin J, Isard M, Zisserman A (2007) Scalable near identical image and shot detection. In: Proceedings of the International Conference on Image and Video Retrieval, pp 549–556

Csurka G, Bray C, Dance C, Fan L (2004) Visual categorization with bags of keypoints. In: Workshop on Statistical Learning in Computer Vision, ECCV, pp 1–22

Dalal N, Triggs B (2005) Histograms of Oriented Gradients for Human Detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 886–893

Demsar J (2006) Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research 7:1–30

Duygulu P, Barnard K, de Freitas N, Forsyth DA (2002) Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In: Proceedings of the European Conference on Computer Vision, pp 97–112

Everingham M, Zisserman A, Williams CKI, Van Gool L (2006a) The 2005 PASCAL visual object classes challenge. In: Machine Learning Challenges – Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Textual Entailment, LNAI, vol 3944, Springer, pp 117–176

Everingham M, Zisserman A, Williams CKI, Van Gool L (2006b) The PASCAL Visual Object Classes challenge 2006 (VOC2006) results. `http://pascal-network.org/challenges/VOC/voc2006/results.pdf`

Everingham M, Van Gool L, Williams CKI, Winn J, Zisserman A (2007) The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. `http://www.pascal-network.org/challenges/VOC/voc2007/index.html`

Fei-Fei L, Fergus R, Perona P (2006) One-shot learning of object categories. IEEE Transactions on Pattern Analysis and Machine Intelligence 28(4):594–611, `http://www.vision.caltech.edu/Image_Datasets/Caltech101/Caltech101.html`

Fellbaum C (ed) (1998) WordNet: an electronic lexical database. MIT Press

Felzenszwalb P, McAllester D, Ramanan D (2008) A discriminatively trained, multiscale, deformable part model. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition

Fergus R, Fei-Fei L, Perona P, Zisserman A (2005) Learning object categories from google's image search. In: Proceedings of the International Conference on Computer Vision

Fergus R, Perona P, Zisserman A (2007) Weakly supervised scale-invariant learning of models for visual recognition. International Journal of Computer Vision 71(3):273–303

Ferrari V, Fevrier L, Jurie F, Schmid C (2008) Groups of adjacent contour segments for object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence 30(1):36–51

Fritz M, Schiele B (2008) Decomposition, discovery and detection of visual categories using topic models. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition

Geusebroek J (2006) Compact object descriptors from local colour invariant histograms. In: Proceedings of the British Machine Vision Conference, pp 1029–1038

Grauman K, Darrell T (2005) The pyramid match kernel: Discriminative classification with sets of image features. In: Proceedings of the International Conference on Computer Vision, pp 1458–1465

Griffin G, Holub A, Perona P (2007) Caltech-256 object category dataset. Tech. Rep. 7694, California Institute of Technology, http://www.vision.caltech.edu/Image_Datasets/Caltech256/

Hoiem D, Efros AA, Hebert M (2006) Putting objects in perspective. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 2137–2144

Kohli P, Ladicky L, Torr P (2008) Robust higher order potentials for enforcing label consistency. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition

Lampert CH, Blaschko MB, Hofmann T (2008) Beyond sliding windows: Object localization by efficient subwindow search. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition

Laptev I (2006) Improvements of object detection using boosted histograms. In: Proceedings of the British Machine Vision Conference, pp 949–958

Lazebnik S, Schmid C, Ponce J (2006) Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 2169–2178

Leibe B, Leonardis A, Schiele B (2004) Combined object categorization and segmentation with an implicit shape model. In: ECCV2004 Workshop on Statistical Learning in Computer Vision, Prague, Czech Republic, pp 17–32

Liu X, Wang D, Li J, Zhang B (2007) The feature and spatial covariant kernel: Adding implicit spatial constraints to histogram. In: Proceedings of the International Conference on Image and Video Retrieval

Lowe D (2004) Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60(2):91–110

Marszalek M, Schmid C (2007) Semantic hierarchies for visual object recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition

Perronnin F, Dance C (2007) Fisher kernels on visual vocabularies for image categorization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition

Pinto N, Cox D, DiCarlo J (2008) Why is real-world visual object recognition hard? PLoS Computational Biology 4(1):151–156

Russell B, Torralba A, Murphy K, Freeman WT (2008) LabelMe: a database and web-based tool for image annotation. International Journal of Computer Vision 77(1–3):157–173, http://labelme.csail.mit.edu/

Salton G, Mcgill MJ (1986) Introduction to Modern Information Retrieval. McGraw-Hill, Inc., New York, NY, USA

Scharstein D, Szeliski R (2002) A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. International Journal of Computer Vision 47(1–3):7–42, http://vision.middlebury.edu/stereo/

Shotton J, Winn JM, Rother C, Criminisi A (2006) *TextonBoost*: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In: Proceedings of the European Conference on Computer Vision, pp 1–15

Sivic J, Zisserman A (2003) Video Google: A text retrieval approach to object matching in videos. In: Proceedings of the International Conference on Computer Vision, vol 2, pp 1470–1477, URL http://www.robots.ox.ac.uk/~vgg

Smeaton AF, Over P, Kraaij W (2006) Evaluation campaigns and TRECVID. In: MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval, pp 321–330

Snoek C, Worring M, Smeulders A (2005) Early versus late fusion in semantic video analysis. In: Proceedings of the ACM International Conference on Multimedia, pp 399–402

Snoek C, Worring M, van Gemert J, Geusebroek J, Smeulders A (2006) The challenge problem for automated detection of 101 semantic concepts in multimedia. In: Proceedings of ACM Multimedia

Sorokin A, Forsyth D (2008) Utility data annotation with amazon mechanical turk. In: Procedings of the First IEEE Workshop on Internet Vision (at CVPR 2008)

Spain M, Perona P (2008) Some objects are more equal than others: Measuring and predicting importance. In: Proceedings of the European Conference on Computer Vision, pp 523–536

Stoettinger J, Hanbury A, Sebe N, Gevers T (2007) Do colour interest points improve image retrieval? In: Proceedings of the IEEE International Conference on Image Processing, pp 169–172

Sudderth EB, Torralba AB, Freeman WT, Willsky AS (2008) Describing visual scenes using transformed objects and parts. International Journal of Computer Vision 77(1–3):291–330

Torralba AB (2003) Contextual priming for object detection. International Journal of Computer Vision 53(2):169–191

Torralba AB, Murphy KP, Freeman WT (2007) Sharing visual features for multiclass and multiview object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence 29(5):854–869

van de Sande KEA, Gevers T, Snoek CGM (2008) Evaluation of color descriptors for object and scene recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition

van de Weijer J, Schmid C (2006) Coloring local feature extraction. In: Proceedings of the European Conference on Computer Vision

van Gemert J, Geusebroek J, Veenman C, Snoek C, Smeulders A (2006) Robust scene categorization by learning image statistics in context. In: CVPR Workshop on Semantic Learning Applications in Multimedia

Viitaniemi V, Laaksonen J (2008) Evaluation of techniques for image classification, object detection and object segmentation. Tech. Rep. TKK-ICS-R2, Department of Information and Computer Science, Helsinki University of Technology, http://www.cis.hut.fi/projects/cbir/

Viola PA, Jones MJ (2004) Robust Real-time Face Detection. International Journal of Computer Vision 57(2):137–154

von Ahn L, Dabbish L (2004) Labeling images with a computer game. In: Proceedings of the ACM CHI, pp 319–326

Wang D, Li J, Zhang B (2006) Relay boost fusion for learning rare concepts in multimedia. In: Proceedings of the International Conference on Image and Video Retrieval

Winn J, Everingham M (2007) The PASCAL Visual Object Classes challenge 2007 (VOC2007) annotation guidelines. http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2007/guidelines.html

Yao B, Yang X, Zhu SC (2007) Introduction to a large scale general purpose ground truth dataset: methodology, annotation tool, and benchmarks. In: Proceedings of the 6th International Conference on Energy Minimization Methods in Computer Vision and Pattern Recognition, http://www.imageparsing.com/

Yilmaz E, Aslam J (2006) Estimating average precision with incomplete and imperfect judgments. In: Fifteenth ACM International Conference on Information and Knowledge Management (CIKM)

Zehnder P, Koller-Meier E, Van Gool L (2008) An efficient multi-class detection cascade. In: Proceedings of the British Machine Vision Conference

Zhang J, Marszalek M, Lazebnik S, Schmid C (2007) Local features and kernels for classification of texture and object categories: A comprehensive study. International Journal of Computer Vision 73(2):213–238