Diagnosing Error in Object Detectors

Derek Hoiem, Yodsawalai Chodpathumwan, and Qieyun Dai *

Department of Computer Science
University of Illinois at Urbana-Champaign

Abstract. This paper shows how to analyze the influences of object characteristics on detection performance and the frequency and impact of different types of false positives. In particular, we examine effects of occlusion, size, aspect ratio, visibility of parts, viewpoint, localization error, and confusion with semantically similar objects, other labeled objects, and background. We analyze two classes of detectors: the Vedaldi et al. multiple kernel learning detector and different versions of the Felzenszwalb et al. detector. Our study shows that sensitivity to size, localization error, and confusion with similar objects are the most impactful forms of error. Our analysis also reveals that many different kinds of improvement are necessary to achieve large gains, making more detailed analysis essential for the progress of recognition research. By making our software and annotations available, we make it effortless for future researchers to perform similar analysis.

1 Introduction

Large datasets are a boon to object recognition, yielding powerful detectors trained on hundreds of examples. Dozens of papers are published every year citing recognition accuracy or average precision results, computed over thousands of images from CalTech or PASCAL VOC datasets [1, 2]. Yet such performance summaries do not tell us *why* one method outperforms another or help understand how it could be improved. For authors, it is difficult to find the most illustrative qualitative results, and the readers may be suspicious of a few hand-picked successes or "intuitive" failures. To make matters worse, a dramatic improvement in one aspect of recognition may produce only a tiny change in overall performance. For example, our study shows that complete robustness to occlusion would lead to an improvement of only a few percent average precision. There are many potential causes of failure: occlusion, intra-class variations, pose, camera position, localization error, and confusion with similar objects or textured backgrounds. To make progress, we need to better understand what most needs improvement and whether a new idea produces the desired effect. We need to measure the modes of failure of our algorithms. Fortunately, we already have excellent large, diverse datasets; now, we propose annotations and analysis tools to take full advantage.

This paper analyzes the influences of object characteristics on detection performance and the frequency and impact of different types of false positives. In particular, we examine effects of occlusion, size, aspect ratio, visibility of parts, viewpoint, localization

* This work was supported by NSF awards IIS-1053768 and IIS-0904209, ONR MURI Grant N000141010934, and a research award from Google.

error, confusion with semantically similar objects, confusion with other labeled objects, and confusion with background. We analyze two types of detectors on the PASCAL VOC 2007 dataset [2]: the Vedaldi et al. [3] multiple kernel learning detector (called VGVZ for authors' initials) and the Felzenszwalb et al. [4, 5] detector (called FGMR). By making our analysis software and annotations available, we will make it effortless for future researchers to perform similar analysis.

Relation to Existing Studies: Many methods have been proposed to address a particular recognition challenge, such as occlusion (e.g., [6–8]), variation in aspect ratio (e.g., [5, 3, 9]), or changes of viewpoint (e.g., [10–12]). Such methods are usually evaluated based on overall performance, home-brewed datasets, or artificial manipulations, making it difficult to determine whether the motivating challenge is addressed for naturally varying examples. Researchers are aware of the value of additional analysis, but there are no standards for performing or compactly summarizing detailed evaluations.

Some studies focus on particular aspects of recognition, such as interest point detection (e.g., [13, 14]), contextual methods [15, 16], dataset design [17], and cross-dataset generalization [18]. The work by Divvala et al. [15] is particularly relevant: through analysis of false positives, they show that context reduces confusion with textured background patches and increases confusion with semantically similar objects. The report on PASCAL VOC 2007 by Everingham et al. [19] is also related in its sensitivity analysis of size and qualitative analysis of failures. For size analysis, Everingham et al. measure the average precision (AP) for increasing size thresholds (smaller-than-threshold objects are treated as "don't cares"). Because the measure is cumulative, some effects of size are obscured, causing the authors to conclude that detectors have a "limited preference for large objects". In contrast, our experiments indicate that object size is the best single predictor of performance.

Studies in specialized domains, such as pedestrian detection [20] and face recognition [21, 22], have provided useful insights. For example, Dollar et al. [20] analyze effects of scale, occlusion, and aspect ratio for a large dataset of pedestrian videos, summarizing performance for different subsets with a single point on the ROC curve.

Contributions: Our main contribution is to provide analysis tools (annotations, software, and techniques) that facilitate detailed and meaningful investigation of object detector performance. Although benchmark metrics, such as AP, are well-established, we had much difficulty to quantitatively explore causes and correlates of error; we wish to share some of the methodology that we found most useful with the community.

Our second contribution is the analysis of two state-of-the-art detectors. We intend our analysis to serve as an example of how other researchers can evaluate the frequency and correlates of error for their own detectors. We sometimes include detailed analysis for the purpose of exemplification, beyond any insights that can be gathered from the results.

Third, our analysis also helps identify significant weaknesses of current approaches and suggests directions for improvement. Equally important, our paper highlights the danger of relying on overall benchmarks to measure short-term progress. Currently, detectors perform well for the most common cases of objects and avoid egregious errors. Large improvements in any one aspect of object recognition may yield small overall gains and go under-appreciated without further analysis.

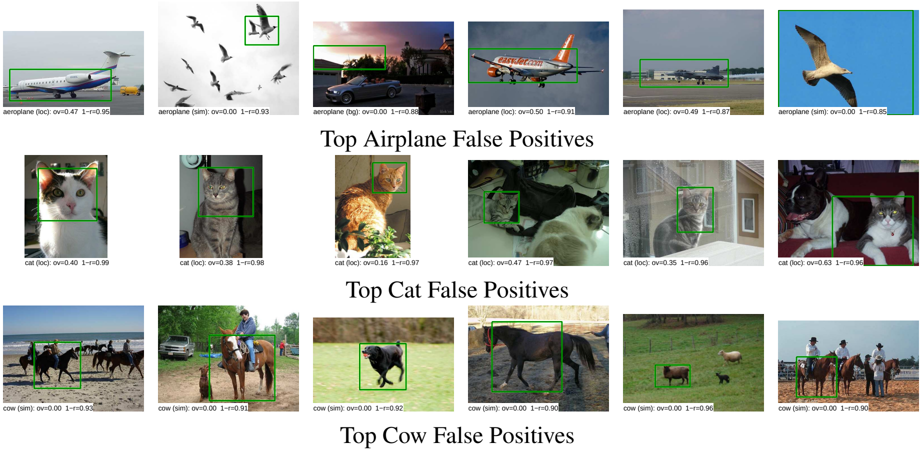
Details of Dataset and Detectors: Our experiments are based on the **PASCAL VOC 2007 dataset** [2], which is widely used to evaluate performance in object category detection. The detection task is to find instances of a specific object category within each input image, localizing each object with a tight bounding box. For the 2007 dataset, roughly 10,000 images were collected from Flickr.com and split evenly into training/validation and testing sets. The dataset is favored for its representative sample of consumer photographs and rigorous annotation and collection procedures. The dataset contains bounding box annotations for 20 object categories, including some auxiliary labels for truncation and five categories of viewpoint. We extend these annotations with more detailed labels for occlusion, part visibility, and viewpoint. Object detections consist of a bounding box and confidence. The highest confidence bounding box with 50% overlap is considered correct; all others are incorrect. Detection performance is measured with precision-recall curves and summarized with average precision. We use the 2007 version of VOC because the test set annotations are available (unlike later versions), enabling analysis on test performance of various detectors. Experiments on later versions are very likely to yield similar conclusions.

The **FGMR detector** [23] consists of a mixture of deformable part models for each object category, where each mixture component has a global template and a set of deformable parts. Both the global template and the deformable parts are represented by HOG features captured at a coarser and finer scale respectively. A hypothesis score contains both a data term (filter responses) and a spatial prior (deformable cost), and the overall score for each root location is computed based on the best placement of parts. Search is performed by scoring the root template at each position and scale. The training of the object model is posed as a latent SVM problem where the latent variables specify the object configuration. The main differences between v4 and v2 are that v4 includes more components (3 vs. 2) and more latent parts (8 vs. 6) and has a latent left/right flip term. We do not use the optional context rescoring.

The **VGVZ detector** [3] adopts a cascade approach by training a three-stage classifier, using as features Bag of Visual Words, Dense Words, Histogram of Oriented Edges and Self-Similarity Features. For each feature channel, a three-level pyramid of spatial histograms is computed. In the first stage, a linear SVM is used to output a set of candidate regions, which are passed to the more powerful second and third stages, which uses quasi-linear and non-linear kernel SVMs respectively. Search is performed using "jumping windows" proposed based on a learned set of detected keypoints.

2 Analysis of False Positives

One major type of error is false positives, detections that do not correspond to the target category. There are different types of false positives which likely require different kinds of solutions. **Localization error** occurs when an object from the target category is detected with a misaligned bounding box ($0.1 <=$ overlap < 0.5). Other overlap


aeroplane (loc): ov=0.47 1−r=0.95
aeroplane (sim): ov=0.00 1−r=0.93
aeroplane (bg): ov=0.00 1−r=0.88
aeroplane (loc): ov=0.50 1−r=0.91
aeroplane (loc): ov=0.49 1−r=0.87
aeroplane (sim): ov=0.00 1−r=0.85
cat (loc): ov=0.40 1−r=0.99
cat (loc): ov=0.38 1−r=0.98
cat (loc): ov=0.16 1−r=0.97
cat (loc): ov=0.47 1−r=0.97
cat (loc): ov=0.35 1−r=0.96
cat (loc): ov=0.63 1−r=0.96
cow (sim): ov=0.00 1−r=0.93
cow (sim): ov=0.00 1−r=0.91
cow (sim): ov=0.00 1−r=0.92
cow (sim): ov=0.00 1−r=0.90
cow (sim): ov=0.00 1−r=0.96
cow (sim): ov=0.00 1−r=0.90

Top Airplane False Positives

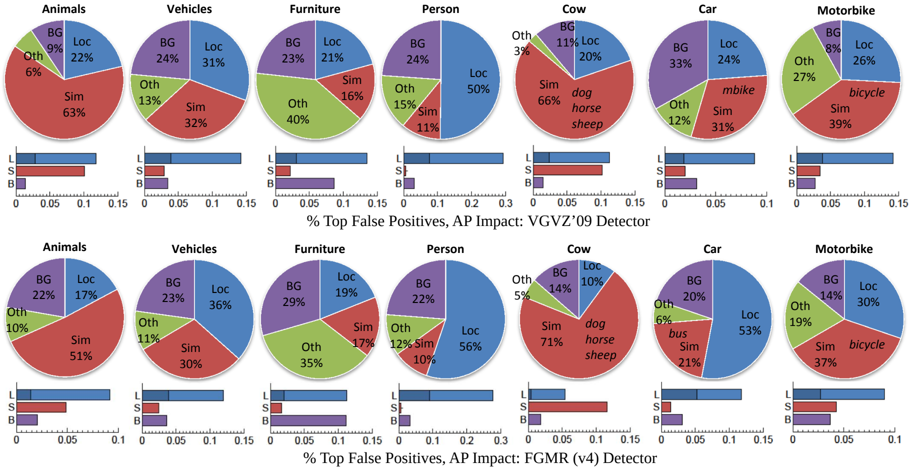
Top Cat False Positives

Top Cow False Positives

Fig. 1. Examples of top false positives: We show the top six false positives (FPs) for the FGMR (v4) airplane, cat, and cow detectors. The text indicates the type of error (“loc”=localization; “bg”=background; “sim”=confusion with similar object), the amount of overlap (“ov”) with a true object, and the fraction of correct examples that are ranked lower than the given false positive (“1-r”, for 1-recall). Localization errors could be insufficient overlap (less than 0.5) or duplicate detections. Qualitative examples, such as these, indicate that confusion with similar objects and localization error are much more frequent causes of false positives than mislabeled background patches (which provide many more opportunities for error).

thresholds (e.g., $0.2 <=$ overlap < 0.5) led to similar conclusions. “Overlap” is defined as the intersection divided by union of the ground truth and detection bounding boxes. We also consider a duplicate detection (two detections for one object) to be localization error because such mistakes are avoidable with good localization. Remaining false positives that have at least 0.1 overlap with an object from a similar category are counted as **confusion with similar objects**. For example a “dog” detector may assign a high score to a “cat” region. We consider two categories to be semantically similar if they are both within one of these sets: {all vehicles}, {all animals including person}, {chair, diningtable, sofa}, {aeroplane, bird}. **Confusion with dissimilar objects** describes remaining false positives that have at least 0.1 overlap with another labeled VOC object. For example, the FGMR bottle detector very frequently detects people because the exterior contours are similar. All other false positives are categorized as **confusion with background**. These could be detections within highly textured areas or confusions with unlabeled objects that are not within the VOC categories. See Fig. 1 for examples of confident false positives.

In Figure 2, we show the frequency and impact on performance of each type of false positive. We count “top-ranked” false positives that are within the most confident N_j detections. We choose parameter N_j to be the number of positive examples for the category, so that if all objects are correctly detected, no false positives would remain. A surprisingly small fraction of confident false positives are due to confusion with background (e.g., only 9% for VGVZ animal detectors on average). For animals, most false positives are due to confusion with other animals; for vehicles, localization and con-



% Top False Positives, AP Impact: VGVZ'09 Detector

% Top False Positives, AP Impact: FGMR (v4) Detector

Fig. 2. Analysis of Top-Ranked False Positives. Pie charts: fraction of top-ranked false positives that are due to poor localization (Loc), confusion with similar objects (Sim), confusion with other VOC objects (Oth), or confusion with background or unlabeled objects (BG). Each category named within 'Sim' is the source of at least 10% of the top false positives. Bar graphs display absolute AP improvement by removing all false positives of one type ('B' removes confusion with background and non-similar objects). 'L': the first bar segment displays improvement if duplicate or poor localizations are removed; the second displays improvement if the localization errors were corrected, turning false detections into true positives.

fusion with similar categories are both common. In looking at trends of false positives with increasing rank, localization errors and confusion with similar objects tend to be more common among the top-ranked than the lower-ranked false positives. Confusion with "other" objects and confusion with background may be similar types of errors. Some categories were often confused with semantically dissimilar categories. For example, bottles were often confused with people, due to the similarity of exterior contours. Removing only one type of false positives may have a small effect, due to the $\frac{TP}{TP+FP}$ form of precision. In particular, the potential improvement by removing all background false detections is surprisingly small (e.g., 0.02 AP for animals, 0.04 AP for vehicles). Improvements in localization or differentiating between similar categories would lead to the largest gains. If poor localizations were corrected, e.g. with an effective category-based segmentation method, performance would improve greatly from additional high-confidence true positives, as well as fewer false positives.

3 False Negatives and Impact of Object Characteristics

Detectors may incur a false negative by assigning a low confidence to an object or by missing it completely. Intuitively, an object may be difficult to detect due to occlusion, truncation, small size, or unusual viewpoint. In this section, we measure the sensitivity of detectors to these characteristics and others and also try to answer why so many objects (typically about 40%) are not detected with even very low confidence.

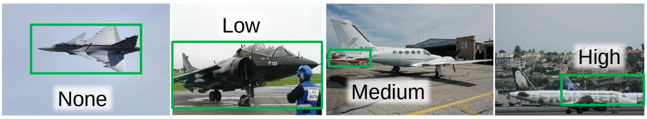


Fig. 3. Example of 4 levels of occlusion for the aeroplane class.

3.1 Definitions of Object Characteristics

To perform our study, we added annotations to the PASCAL VOC dataset, including level of occlusion and which sides and parts of an object are visible. We also use standard annotations, such as the bounding box. We created the extra annotations for seven categories (‘aeroplane’, ‘bicycle’, ‘bird’, ‘boat’, ‘cat’, ‘chair’, ‘diningtable’) that span the major groups of vehicles, animals, and furniture. Annotations were created by one author to ensure consistency and are publicly available.

Object size is measured as the pixel area of the bounding box. We also considered bounding box height as a size measure, which led to similar conclusions. We assign each object to a size category, depending on the object’s percentile size within its object category: extra-small (XS: bottom 10%); small (S: next 20%); medium (M: next 40%); large (L: next 20%); extra-large (XL: next 10%). **Aspect ratio** is defined as object width divided by object height, computed from the VOC bounding box annotation. Similarly to object size, objects are categorized into extra-tall (XT), tall (T), medium (M), wide (W), and extra-wide (XW), using the same percentiles. **Occlusion** (part of the object is obscured by another surface) and **truncation** (part of the object is outside the image) have binary annotations in the standard VOC annotation. We replace the occlusion labels with degree of occlusion (see Fig. 3): ‘None’, ‘Low’ (slight occlusion), ‘Moderate’ (significant part is occluded), and ‘High’ (many parts missing or 75% occluded). **Visibility of parts** influences detector performance, so we add annotations for whether each part of an object is visible. We annotate **viewpoint** as whether each side (‘bottom’, ‘front’, ‘top’, ‘side’, ‘rear’) is visible. For example, an object seen from the front-right may be labeled as ‘bottom’=0, ‘front’=1, ‘top’=0, ‘side’=1, ‘rear’=0.

3.2 Normalized Precision Measure

To analyze sensitivity to object characteristics, we would like to summarize the performance of different subsets (e.g. small vs. large) of objects. Current performance measures do not suffice: ROC curves are difficult to summarize, and average precision is sensitive to the number of positive examples. We propose a simple way to normalize precision so that we can easily measure and compare performance for objects with particular characteristics.

The standard PASCAL VOC measure is average precision (AP), which summarizes precision-recall curves with the average interpolated precision value of the positive examples. Recall $R(c)$ is the fraction of objects detected with confidence of at least c.

		aero	bike	boat	bus	car	mbike	train	bird	cat	cow	dog	horse	sheep	bottle	chair	table	plant	sofa	tv	pers
Num Objs		285	337	263	213	1201	325	282	459	358	244	489	348	242	469	756	206	480	239	308	4528
VGVZ	AP	.364	.468	.113	.513	.508	.450	.447	.106	.277	.312	.174	.512	.208	.193	.131	.191	.073	.266	.477	.200
	AP$_{\mathrm{N}}$	.443	.531	.181	.612	.480	.511	.527	.136	.372	.418	.222	.565	.294	.222	.130	.318	.094	.294	.567	.073
FGMR	AP	.277	.581	.134	.481	.555	.475	.436	.028	.149	.214	.034	.582	.154	.225	.191	.204	.068	.306	.405	.410
	AP$_{\mathrm{N}}$	.343	.631	.187	.574	.517	0.537	.528	.039	.207	.311	.048	.631	.215	.252	.190	.346	.086	.409	.477	.214

Table 1. Detection Results on PASCAL VOC2007 Dataset. For each object category, we show the total number of objects and the average precision (AP) and average normalized precision (AP$_{\mathrm{N}}$) for the VGVZ and FGMR detectors. For AP$_{\mathrm{N}}$, the precision is normalized to be comparable across categories. Categories with many positive examples (e.g., "person") will have lower AP$_{\mathrm{N}}$ than AP; the reverse is true for categories with few examples.

Precision $P(c)$ is the fraction of detections that are correct:

$$P(c) = \frac{R(c) \cdot N_j}{R(c) \cdot N_j + F(c)} \tag{1}$$

where N_j is the number of objects in class j and $F(c)$ is the number of incorrect detections with at least c confidence. Before computing AP, precision is "interpolated", such that the interpolated precision value at c is the maximum precision value for any example with at least confidence c. If N_j is large (such as for pedestrians), then the precision would be higher than if N_j were small for the same detection rate. This sensitivity to N_j invalidates AP comparisons for different sets of objects. For example, we cannot use AP to determine whether people are easier to detect than cows, or whether big objects are easier to detect than small ones.

We propose to replace N_j with a constant N to create a normalized precision:

$$P_{\mathrm{N}}(c) = \frac{R(c) \cdot N}{R(c) \cdot N + F(c)}. \tag{2}$$

In our experiments, we set $N = 0.15 N_{images} = 742.8$, which is roughly equal to the average N_j over the PASCAL VOC categories. Detectors with similar detection rates and false positive rates will have similar normalized precision-recall curves. We can summarize normalized precision-recall by averaging the normalized precision values of the positive examples to create average normalized precision (AP$_{\mathrm{N}}$). When computing AP$_{\mathrm{N}}$, undetected objects are assigned a precision value of 0. We compare AP and AP$_{\mathrm{N}}$ in Table 1.

3.3 Analysis

To understand performance for a category, it helps to inspect the performance variations for each characteristic. We include this detailed analysis primarily as an example of how researchers can use our tool to understand the strengths and weaknesses of their own detectors. Upon careful inspection of Figure 4, for example, we can learn the following about airplane detectors: both detectors perform similarly, preferring similar subsets of non-occluded, non-truncated, medium to extra-wide, side views in which all major parts are visible. Performance for very small and heavily occluded airplanes is poor,

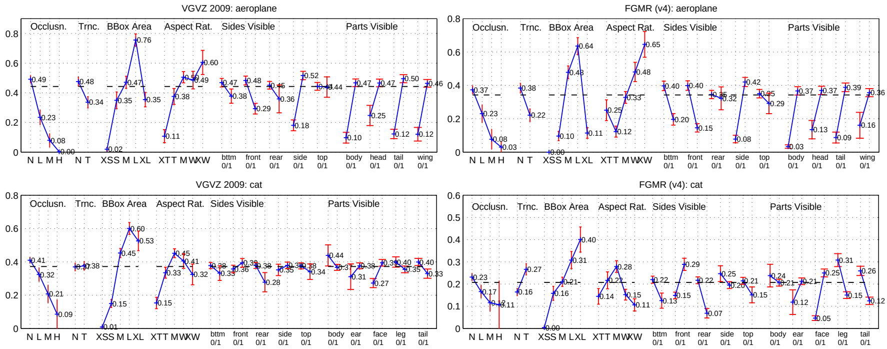


Fig. 4. Per-Category Analysis of Characteristics: AP_N ('+') with standard error bars (red). Black dashed lines indicate overall AP_N. Key: *Occlusion*: N=none; L=low; M=medium; H=high. *Truncation*: N=not truncated; T=truncated. *Bounding Box Area*: XS=extra-small; S=small; M=medium; L=large; XL =extra-large. *Aspect Ratio*: XT=extra-tall/narrow; T=tall; M=medium; W=wide; XW =extra-wide. *Part Visibility / Viewpoint*: '1'=part/side is visible; '0'=part/side is not visible. Standard error is used for the average precision statistic as a measure of significance, rather than confidence bounds, due to the difficulty of modeling the precision distribution.

though performance on non-occluded airplanes is near average (because few airplanes are heavily occluded). FGMR shows a stronger preference for exact side-views than VGVZ, which may also account for differences in performance on small and extra-large objects, which are both less likely to be side views. We can learn similar things about the other categories. For example, both detectors work best for large cats, and FGMR performs best with truncated cats, where all parts except the face and ears are not visible. Note that both detectors seem to vary in similar ways, indicating that their sensitivities may be due to some objects being intrinsically more difficult to recognize. The VGVZ cat detector is less sensitive to viewpoint and part visibility, which may be due to its textural bag of words features.

Since size and occlusion are such important characteristics, we think it worthwhile to examine their effects across several categories (Fig. 5). Typically, detectors work best for non-occluded objects, but when objects are frequently occluded (bicycle, chair, diningtable) there is sometimes a small gain in performance for lightly occluded objects. Although detectors are bad at detecting medium-heavy occluded objects, the impact on overall performance is small. Researchers working on the important problem of occlusion robustness should be careful to examine effects within the subset of occluded objects. Both detectors tend to prefer medium to large objects (the 30th to 90th percentile in area). The difficulty with small objects is intuitive. The difficulty with extra-large objects may initially surprise, but qualitative analysis (e.g., Fig. 7) shows that large objects are often highly truncated or have unusual viewpoints.

Fig. 6 provides a compact summary of the sensitivity to each characteristic and the potential impact of improving robustness. The worst-performing and best-performing subsets for each characteristic are averaged over 7 categories: the difference between best and worst indicates sensitivity; the difference between best and overall indicates

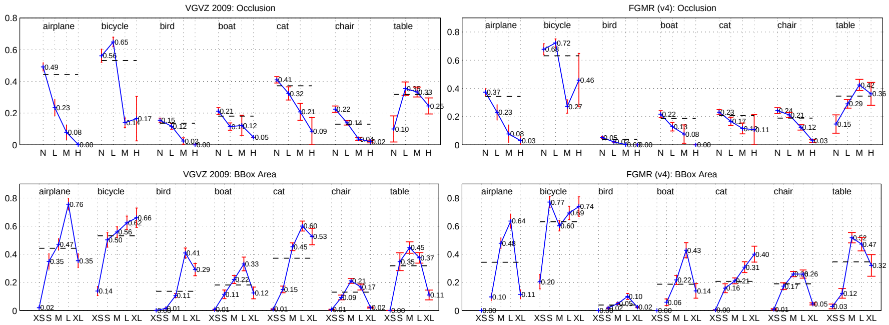


Fig. 5. Sensitivity and Impact of Object Characteristics: AP_N ('+') with standard error bars (red). Black dashed lines indicate overall AP_N. Key: *Occlusion*: N=none; L=low; M=medium; H=high. *Bounding Box Area*: XS=extra-small; S=small; M=medium; L=large; XL =extra-large.

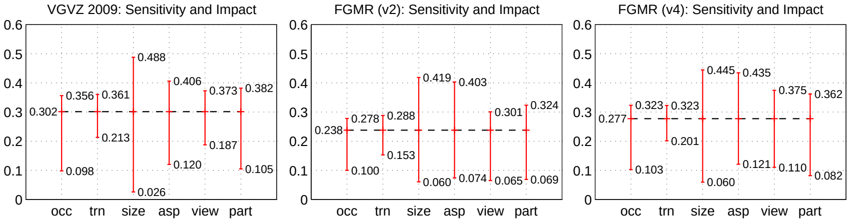


Fig. 6. Summary of Sensitivity and Impact of Object Characteristics: We show the average (over categories) AP_N performance of the highest performing and lowest performing subsets within each characteristic (occlusion, truncation, bounding box area, aspect ratio, viewpoint, part visibility). Overall AP_N is indicated by the black dashed line. The difference between max and min indicates sensitivity; the difference between max and overall indicates the impact.

potential impact. For example, detectors are very sensitive to occlusion and truncation, but the impact is small (about 0.05 AP_N) because the most difficult cases are not common. Object area and aspect have a much larger impact (roughly 0.18 for area, 0.13 for aspect). Overall, VGVZ is more robust than FGMR to viewpoint and part visibility, likely because it is better at encoding texture (due to bags of words features), while FGMR can accommodate only limited deformations. The performance of VGVZ varies more than FGMR with object size. Efforts to improve occlusion or viewpoint robustness should be validated with specialized analysis so that improvements are not diluted by the more common easier cases.

One of the aims of our study is to understand why current detectors fail to detect $30 - 40\%$ of objects, even at small confidence thresholds. We consider an object to be undetected if there is no detection above 0.05 P_N (roughly 1.5 FP/image). Our analysis indicates that size is the best single explanation, as nearly all extra-small objects go undetected and roughly half of undetected objects are in the smallest 30%. But size accounts for only 20% of the entropy of whether an object is detected (measured by information gain divided by entropy). Contrary to intuition, occlusion does not increase the chance of going undetected (though it decreases the expected confidence); there is

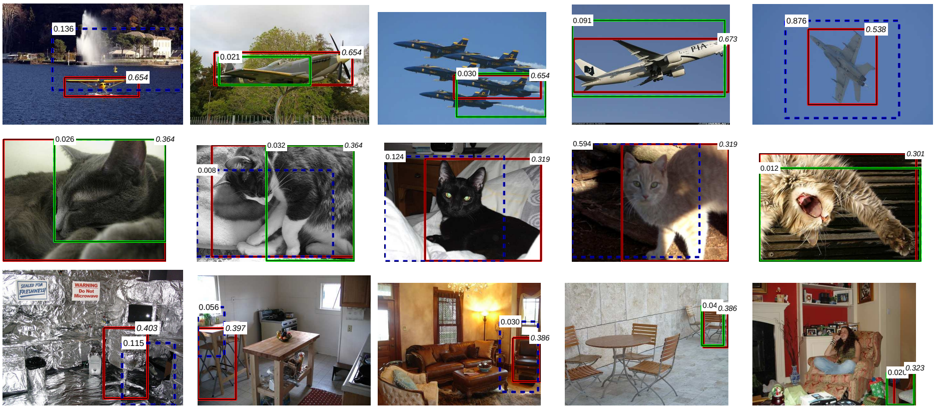


Fig. 7. Unexpectedly Difficult Detections: We fit a linear regressor to predict confidence (P_N) based on size, aspect ratio, occlusion and truncation for the FGMR (v4) detector. We show 5 of the 15 objects with the greatest difference between predicted confidence and actual detection confidence. The ground truth object is in red, with predicted confidence in upper-right corner in italics. The other box shows the highest scoring correct detection (green), if any, or highest scoring overlapping detection (blue dashed) with the detection confidence in the upper-left corner.

also only a weak correlation with aspect ratio. As can be seen in Figure 7, misses are due to a variety of factors, such as unusual appearance, unusual viewpoints, in-plane rotation, and particularly disguising occlusions. Some missed detections are actually detected with high confidence but poor localization. With a 10% overlap criteria, the number of missed detections is reduced by 40-85% (depending on category, detector).

Conclusions and Discussion: As expected, objects that are small, heavily occluded, seen from unusual views (e.g., the bottom), or that have important occluded parts are hard to detect. But the deviations are interesting: slightly occluded bicycles and tables are easier; detectors often have trouble with very large objects; cats are easiest for FGMR when only the head is visible. Despite high sensitivity, the overall impact of occlusion and many other characteristics is surprisingly small. Even if the gap between no occlusions and heavy occlusions were completely closed, the overall AP gain would be only a few percent. Also, note that smaller objects, heavily occluded objects, and those seen from unusual viewpoints will be intrinsically more difficult, so we cannot expect to close the gap completely.

We can also see the impact of design differences for two detectors. The VGVZ detector incorporates more textural features and also has a more flexible window proposal method. These detector properties are advantageous for detecting highly deformable objects, such as cats, and they lead to reduced sensitivity to part visibility and viewpoint. However, VGVZ is more sensitive to size (performing better for large objects and worse for small ones) because the bag-of-words models are size-dependent, while the FGMR templates are not. Comparing the FGMR (v2) and (v4) detectors in Figure 6, we see that the additional components and left/right latent flip term lead to improved robustness (improvement in both worst and best cases) to aspect ratio, viewpoint, and truncation.

4 Conclusion

4.1 Diagnosis

We have analyzed the patterns of false and missed detections made by currently top-performing detectors. The good news is that egregious errors are rare. Most false positives are due to misaligned detection windows or confusion with similar objects (Fig. 2). Most missed detections are atypical in some way – small, occluded, viewed from an unusual angle, or simply odd looking. Detectors seem to excel at latching onto the common modes of object appearance and avoiding confusion with miscellaneous background patches. For example, detectors more easily detect lightly occluded objects for categories that are typically occluded, such as bicycles and tables. Airplanes in the stereotypical direct side-view are detected by FGMR with almost double the accuracy of the average airplane.

Further progress, however, is likely to be slow, incremental, and to require careful validation. Localization error is not easily handled by detectors because determining object extent often requires looking well outside the window. A box around a cat face could be a perfect detection in one image but only a small part of the visible cat in another (Figs. 1, 8). Gradient histogram-based models, designed to accommodate moderate positional slop, may be too coarse for differentiation of similar objects, such as cows and horses or cats and dogs. Although detectors fare well for common views of objects, there are many types of deviations that may require different solutions. Solving only one problem will lead to small gains. For example, even if occluded objects could be detected as easily as unoccluded ones, the overall AP would increase by 0.05 or less (Fig. 6). Better detection of small objects could yield large gains (roughly 0.17 AP), but the smallest objects are intrinsically difficult to detect. Our biggest concern is that successful attempts to address problems of occlusion, size, localization, or other problems could look unpromising if viewed only through the lens of overall performance. There may be a gridlock in recognition research: new approaches that do not immediately excel at detecting objects typical of the standard datasets may be discarded before they are fully developed. One way to escape that gridlock is through more targeted efforts that specifically evaluate gains along the various dimensions of error.

4.2 Recommendations

Detecting Small/Large Objects: An object's size is a good predictor of whether it will be detected. Small objects have fewer visual features; large objects often have unusual viewpoints or are truncated. Current detectors make a weak perspective assumption that an object's appearance does not depend on its distance to the camera. This assumption is violated when objects are close, making the largest objects difficult to detect. When objects are far away, only low resolution template models can be applied, and contextual models may be necessary for disambiguation. There are interesting challenges in how to benefit from the high resolution of nearby objects while being robust to near-field perspective effects. Park et al. [24] offer one simple approach of combining detectors

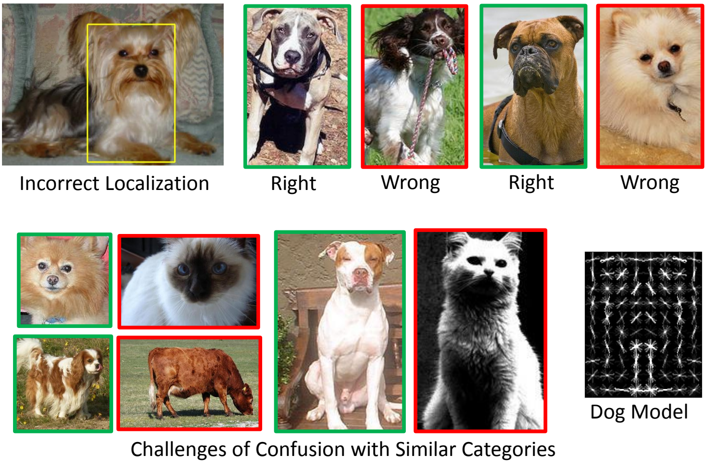


Fig. 8. Challenging False Positives: In the top row, we show examples of dog detections that are considered correct or incorrect, according to the VOC localization criteria. On the right, we cropped out the rest of the image; it is not possible to determine correctness from within the window alone. On the bottom, we show several confident dog detections, some of which correspond to objects from other similar categories. The robust HOG template detector (right), though good for sorting through many background patches quickly, may be too robust for more fine differentiations.

trained at different resolutions. Other possibilities include using scene layout to predict viewpoint of nearby objects, including features that take advantage of texture and small parts that are visible for close objects, and occlusion reasoning for distant objects.

Improving Localization: Difficulty with localization and identification of duplicate detections has a large impact on recognition performance for many categories. In part, the problem is that an template-based detectors cannot accommodate flexible objects. For example, the FGMR (v4) detector works very well for localizing cat heads, but not cat bodies [25]. Additionally, an object part, such as a head, could be considered correct if the body is occluded, or incorrect otherwise; it is impossible to tell from within the window alone (Fig. 8). Template detectors should play a major role in finding likely object positions, but specialized processes are required for segmenting out the objects. In some applications, precise localization is unimportant; even so, the ability to identify occluding and interior object contours would also be useful for category verification, attribute recognition, or pose estimation. Current approaches to segment objects from known categories (e.g., [8, 25]) tend to combine simple shape and/or color priors with more generic segmentation techniques, such as graph cuts. There is an excellent opportunity for a more careful exploration of the interaction of material, object shape, pose, and object category, through contour and texture-based inference. We encourage such work to evaluate specifically in terms of improved localization of objects and to avoid conflating detection and localization performance.

Reducing Confusion with Similar Categories: By far, the biggest task of detectors is to quickly discard random background patches, and they do well with features that are made robust to illumination, small shifts, and other variations. But a detector that sorts through millions of windows per second may not be suited for differentiating between dogs and cats or horses and cows. Such differences require detailed comparison of particular features, such as the shape of the head or eyes. Some recent work has addressed fine-grained differentiation of birds and plants [26–28], and the ideas of finding important regions for comparison may apply to category recognition as well. Also, while HOG features [29] are well-suited to whole-object detection, some of the feature representations originally developed for detection of small faces (e.g., [30]) may be better for differentiating similar categories based on their localized parts. We encourage such work to evaluate specifically on differentiating between similar objects. It may be worthwhile to pose the problem simply as categorizing a set of well-localized objects (even categorizing objects given PASCAL VOC bounding boxes is not easy).

Robustness to Object Variation: One interpretation of our results is that existing detectors do well on the most common modes of appearance (e.g., side views of airplanes) but fail when a characteristic, such as viewpoint, is unusual. Greatly improved recognition will require appearance models that better encode shape with robustness to moderate changes in viewpoint, pose, lighting, and texture. One approach is to learn feature representations from paired 3D and RGB images; a second is to learn the natural variations of existing features within categories for which examples are plentiful and to extend that variational knowledge to other categories.

More Detailed Analysis: Most important, we hope that our work will inspire research that targets and evaluates reduction in specific modes of error. In the supplemental material, we include automatically generated reports for four detectors. Authors of future papers can use our tools to perform similar analysis, and the results can be compactly summarized in $1/4$ page, as shown in Figs. 2, 6. We also encourage analysis of other aspects of recognition, such as the effects of training sample size, cross-dataset generalization, cross-category generalization, and recognition of pose and other properties.

References

1. Griffin, G., Holub, A., Perona, P.: Caltech-256 object category dataset. Technical Report 7694, California Institute of Technology (2007)
2. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. (http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html)
3. Vedaldi, A., Gulshan, V., Varma, M., Zisserman, A.: Multiple kernels for object detection. In: ICCV. (2009)
4. Felzenszwalb, P., McAllester, D., Ramanan, D.: A discriminatively trained, multiscale, deformable part model. In: CVPR. (2008)
5. Felzenszwalb, P., Girshick, R., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part based models. PAMI (2009)
6. Wu, B., Nevatia, R.: Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors. In: ICCV. (2005)

7. Wang, X., Han, T.X., Yan, S.: An hog-lbp human detector with partial occlusion handling. In: ICCV. (2009)
8. Yang, Y., Hallman, S., Ramanan, D., Fowlkes, C.: Layered object detection for multi-class segmentation. In: CVPR. (2010)
9. Vedaldi, A., Zisserman, A.: Structured output regression for detection with partial occulsion. In: NIPS. (2009)
10. Kushal, A., Schmid, C., Ponce, J.: Flexible object models for category-level 3d object recognition. In: CVPR. (2007)
11. Hoiem, D., Rother, C., Winn, J.: 3d layoutcrf for multi-view object class recognition and segmentation. In: CVPR. (2007)
12. Sun, M., Su, H., Savarese, S., Fei-Fei, L.: A multi-view probabilistic model for 3d object classes. In: CVPR. (2009)
13. Schmid, C., Mohr, R., Bauckhage, C.: Evaluation of interest point detectors. IJCV **37** (2000) 151–172
14. Gil, A., Mozos, O.M., Ballesta, M., Reinoso, O.: A comparative evaluation of interest point detectors and local descriptors for visual slam. Machine Vision and Applications **21** (2009) 905–920
15. Divvala, S., Hoiem, D., Hays, J., Efros, A., Hebert, M.: An empirical study of context in object detection. In: CVPR. (2009)
16. Rabinovich, A., Belongie, S.: Scenes vs. objects: a comparative study of two approaches to context based recognition. In: Intl. Wkshp. on Visual Scene Understanding (ViSU). (2009)
17. Pinto, N., Cox, D.D., DiCarlo, J.J.: Why is real-world visual object recognition hard? PLoS Computational Biology **4** (2008) e27
18. Torralba, A., Efros, A.: Unbiased look at dataset bias. In: CVPR. (2011)
19. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. IJCV **88** (2010) 303–338
20. Dollar, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian detection: A benchmark. In: CVPR. (2009)
21. Sim, T., Baker, S., Bsat, M.: The CMU Pose, Illumination, and Expression (PIE) database of human faces. Technical Report CMU-RI-TR-01-02, Carnegie Mellon, Robotics Institute (2001)
22. Phillips, P.J., Flynn, P.J., Scruggs, T., Bowyer, K.W., Chang, J., Hoffman, K., Marques, J., Min, J., Worek, W.: Overview of the face recognition grand challenge. (2005) 947–954
23. Felzenszwalb, P.F., Girshick, R.B., McAllester, D.: Discriminatively trained deformable part models, release 4. (http://people.cs.uchicago.edu/ pff/latent-release4/)
24. Park, D., Ramanan, D., Fowlkes, C.: Multiresolution models for object detection. In: ECCV. (2010)
25. Parkhi, O., Vedaldi, A., Jawahar, C.V., Zisserman, A.: The truth about cats and dogs. In: ICCV. (2011)
26. Shirdhonkar, S., White, S., Feiner, S., Jacobs, D., Kress, J., Belhumeur, P.N.: Searching the worlds herbaria: A system for the visual identification of plant species. In: ECCV. (2008)
27. Welinder, P., Branson, S., Mita, T., Wah, C., Schroff, F., Belongie, S., Perona, P.: Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology (2010)
28. Khosla, A., Yao, B., Fei-Fei, L.: Combining randomization and discrimination for fine-grained image categorization. In: CVPR. (2011)
29. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR. (2005)
30. Schneiderman, H., Kanade, T.: A statistical model for 3-d object detection applied to faces and cars. In: CVPR. (2000)