

Visual Recognition, Circa 2007

Pietro Perona
California Institute of Technology

December, 2007 *

Abstract

I am collecting here a few thoughts on broad issues in visual recognition. I am assuming that the reader has some familiarity with vision and visual recognition. This is not a survey, and the references are meant to exemplify an idea or an approach – they are not meant to give proper credit to the many good people who work in the field. Also, some of the interesting issues are not visible from the mile-high perspective I take here, and are therefore not mentioned.

1 Why?

Why worry about visual recognition? I can see two reasons, each one compelling in its own right. The first is that we wish to understand how visual recognition works in biological systems. It is one of the wonders of nature that we can see at all. Understanding vision in computational terms brings us closer to understanding how the brain works. The second reason is that machines whose visual system approaches human ability would be extremely useful in a great number of applications: building human-machine interfaces, searching and indexing into image and video collections, security, manufacturing, monitoring the environment. If we could build such machines our lives would be better¹.

It is useful to reflect on a few of these applications. Understanding their characteristics will help us understand what form of ‘visual recognition’ are potentially most useful, and how to measure a system’s performance. I will focus here on two significant applications that will help me reason about the complexity of visual recognition systems.

The first one is analyzing and indexing large collections of photographs and video. Cameras and video sensors are becoming cheaper and better by the year. Storage space is

***Appeared in: Object Categorization Computer and Human Vision Perspectives. Cambridge University Press 2009. pp. 55-68. Online ISBN: 9780511635465. Hardback ISBN: 9780521887380.**

¹We should not, however, forget that these machines could be used in ways that limit our privacy, as well as to build dangerous weapons.

similarly becoming abundant and inexpensive. As a consequence, large amounts of images are captured and are stored on hard drives; vastly more images than can be inspected and organized by humans. Images and video are quickly becoming a sort of ‘dark matter’: taking 99% of the storage space reserved for data, and being virtually inaccessible. Automating the process of associating keywords with images, linking meaningful visual patches in images with other patches and with text, discovering interesting content in large collections of images would help make these image collections useful. The complexity of image collection may vary widely. An astronomical survey of a section of the sky might contain few, possibly fewer than one hundred, distinct categories. The collection of movies owned by a major film house likely contains tens of thousands, as we shall see later.

A second interesting and useful application is autonomous vehicles driving in urban/suburban traffic. How much recognition is needed? One could reasonably take two opposite and extreme points of view. The ‘minimalistic’ point of view says that no recognition is needed at all: all that a vehicle needs to know is the 3D shape of its immediate environment, and this information will be sufficient for successful navigation, if put in register (e.g. using GPS and inertial sensors) with readily available street maps. At the other extremum, one could argue that for human-like driving performance one must compute any information that is useful at predicting the behavior of pedestrians, vehicles, animals and other obstacles. According to this second point of view, recognition of thousands of categories is vital.

2 Tasks of visual recognition

I distinguish five recognition tasks. In order of difficulty:

Verification - An area of the image has been selected. This may contain a given object: is it there or not? The automated concierge at the entrance of a building is an example of this: a person stands in front of the camera and punches Sally Jean’s PIN. Do I see Sally Jean’s face in the picture? Yes or no?

Detection and localization - A complex image is presented. It may contain an exemplar of a given category. Is it there? Where? Finding human faces in pictures we took during our last vacation is an example of this.

Classification - Does a given image patch contain an object? Which one is it, amongst many possible categories? An example of this is classifying all the faces found by a face detector (see ‘detection’ above) in the pictures I have on my hard drive: 353 pictures of my first child, 233 of my second child, 3 of uncle Joe.

Naming - Given a complex image, name and locate all objects that are present there, out of a (possibly large) number of categories. If we wanted to associate automatically keywords to images, for example to allow word-based indexing without requiring humans to do it by hand, we would want to automate naming.

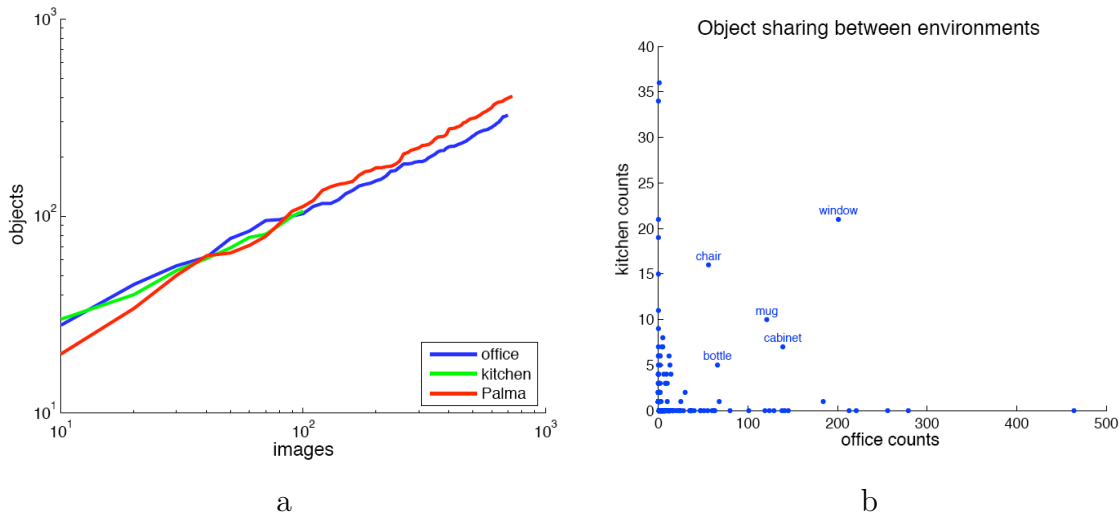


Figure 1: How many things do we recognize? (a) The total number of words reported by human observers when looking at pictures taken in three environments: offices, kitchens and Palma de Mallorca. The number of words increases proportionally to the square root of the number of pictures. There is no sign of saturation even when the number of pictures approaches 10^3 . This suggests that, even in fairly restricted environments, there is a large number of different things that may be recognized. (b) Different environments share few recognizable things. Only ‘bottle’, ‘chair’, ‘cabinet’, ‘mug’ and ‘window’ are shared between kitchens and offices. (Adapted from [Spain and Perona, 2007])

Description - Given an image, describe it: what is the environment, what are the objects and actions you see, what are their relationships. A typical example would be preparing text summaries for a collection of images, so that it may be searched using complex queries: “Find all pictures showing the pope kissing a child.” This is also called ‘scene understanding’.

Which ones of these tasks have been attempted so far? Most ongoing work in visual recognition is in the second category, classification. For instance, Caltech 101 and Caltech 256 images typically contain a single object and little clutter. The task is than one of classifying the entire image as belonging to a given category. It is intuitive that, as soon as we have solved classification, we can move on to naming by shifting windows of different sizes across the image and classifying each such region of interest. Similarly, one could think of solving detection and localization by running a verification algorithm on each window of the image. Unfortunately, this intuition is deceiving, as I will argue in section 4.

3 How many? How fast?

What is the size of the problem we are trying to solve? If our goal is to emulate, and perhaps surpass, human abilities, then we should gear up for 10^5 categories. Biederman estimates that there are $3 \cdot 10^3$ entry-level categories and $3 \cdot 10^4$ visual categories overall (see also

Fig. 1). These numbers make sense: a learned Chinese scholar can recognize $3 \cdot 10^4$ different characters, but a regular person recognizes around $5 \cdot 10^3$. If we recognize similarly a few thousand categories in 10-20 different domains (city streets and outdoor scenes, people, foods, indoor scenes, biology, animals, ...) we end up with a similar estimate: between 10^4 and 10^5 . Arguably, an automatic system could integrate the visual knowledge of many people, each one an expert in a different domain: medicine, paleontology, car mechanics, botany etc, and thus might recognize far more categories than any human would: thus the 10^5 target.

Is 'emulating human abilities' the right way to formulate the problem? After all, we are building machines that are designed for specific environments, while humans operate across many different environments. Thus, a machine's ability to recognize may not need to approach human ability. Going back to the example of autonomous driving, an automobile will operate in cities, suburbs, open landscape and freeways, but it will never enter a kitchen and never look at fMRI images of my brain. Furthermore, a precise classification of many objects that might be visible from a vehicle (e.g. an ice-cream parlor, a larch tree), might well be irrelevant for the task of driving safely down the road. We do not yet have good estimates of the number of categories that are relevant for well-defined tasks. The lower bound might be 10-20 (pedestrians, 3-4 types of vehicles, traffic signs, road markings). The upper bound might reach 10^3 : a good driver will recognize a toy ball, which will tell him that an oblivious child may be running after it. Also: it may be important to distinguish between a small cardboard box, a cinder-block, a crow and a cat; each one of these similarly-sized obstacles might be seen in the middle of the road and each will behave differently when the car approaches and will inflict different amounts of damage to the vehicle if run over.

However: a general-purpose machine for recognition also seems useful: think of classifying by content all images and video one finds on the web. So: it is likely that we will need to design different machines that are optimized for a broad range of task complexities: from 1 to, say, 10^5 categories.

How fast should the task be accomplished? In some cases time is of the essence: think of the autonomous driving vehicle. So: it would be nice to be able to detect and locate 10-100 categories in real-time. What about annotating and organizing large image collections? At first blush, one could convince oneself that a large batch process taking days or weeks to process a few thousands of images (e.g. organizing the pictures I have on my hard drive) would be good enough. However: we must remember that some collections of images are indeed extremely large, are updated frequently, computation sometime comes at a premium, and fast turnaround is also important. Think of a planetary orbiter with a tenuous downlink, collecting thousands of high-resolution pictures every day and having to prioritize which images to send to earth. Think of a TV soap opera, where each episode needs to be annotated and put on line as soon as it is aired.

In conclusion: while visual recognition is already useful for a number of applications, current classification speeds, a few pictures per minute when classifying 1 out of 10^2 categories, are not good enough for most uses. If we keep in mind It would be highly desirable to achieve naming of, say 10^3 categories on many frames per second. We are many orders of

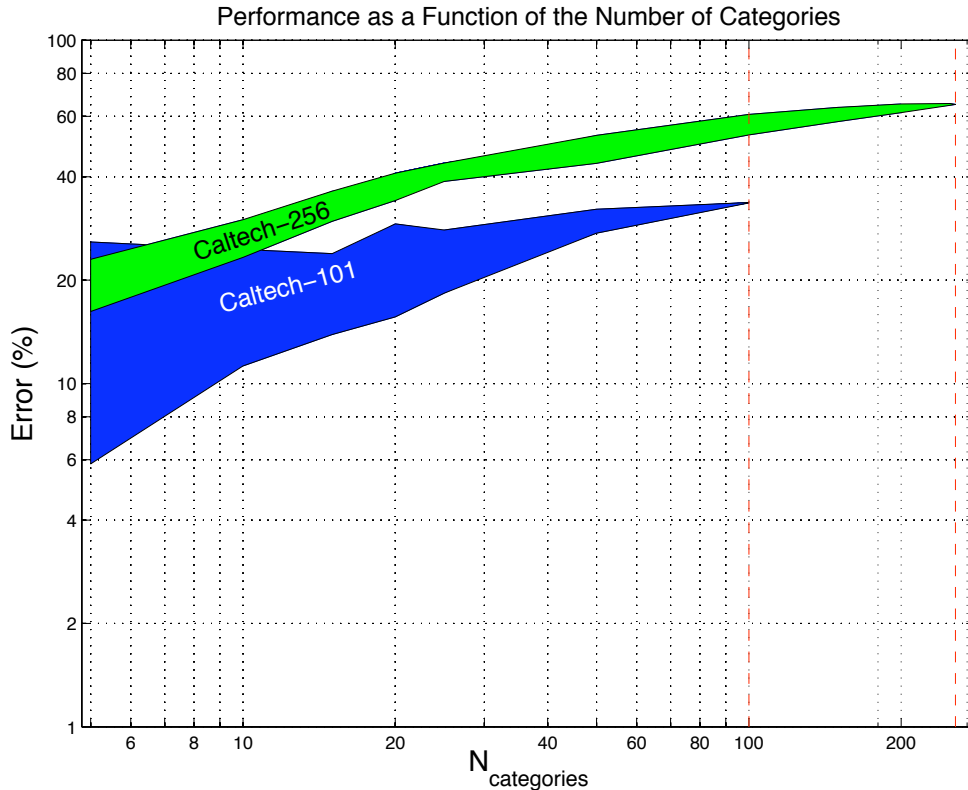


Figure 2: Classification error increases quickly with the number of categories. (Adapted from [Griffin et al., 2007], 30 training examples.)

magnitudes away from achieving this goal.

4 Scaling

One could think of building more complex tasks by combining simpler ones. For example, classification could be obtained by combining the output of multiple verification modules, one per category. Furthermore, detection could be built out of verification: consider all rectangular windows in an image (sampling scale and position) and run verification on each. Will this work? Unfortunately there are a number of difficulties. First, good verification performance in any one category does not predict good classification performance over thousands of categories: the more choices there are, the easier it is to make mistakes. Performance drops precipitously with the number of categories (see Fig. 2. Second, in a straight-out implementation the computational cost would be horrendous: linear both in the number of categories and in the number of test windows, which depending on the sampling scheme, could be 10 x the number of pixels. Third, the probability of any category being present in any given window is small (e.g. $10^{-7} - 10^{-8}$ in a typical scenario), thus, one would be forced

to choose high detection thresholds in order not to be swamped by false alarms, and this would likely decrease detection rates to unacceptable levels. In sum, the answer is “no”, verification does not scale easily to detection, and classification does not scale easily to naming. New ideas are necessary [Viola and Jones, 2004, Fleuret and Geman, 2001].

5 The state of the art

Are we close to having ‘solved’ any of the tasks which are defined above? How many have we tackled credibly?

Most ongoing work is focussed on *classification*. The typical benchmark datasets are Caltech 101 and Caltech 256 [Griffin et al., 2007]. Images from Caltech 101 and Caltech 256 were culled from the web using Google images and other equivalent search engines. Most of these images contain only one object, and the object is rather central in the picture and the viewpoint is conventional [Griffin et al., 2007]. Most algorithms which are tested on Caltech 101 and Caltech 256 are designed to classify the entire image, without separating foreground from background and therefore do not generalize to the task of *naming* in large images where the position of the object is unknown. This said, classification is improving steadily as one may clearly see from Fig. 3.a. How close are we to good performance in classification? It is apparent that there is much year-on-year progress. However, it is clear from Fig. 3.b that we are far from human-level performance (it is reasonable to assume 1% error rate for comparison, although an accurate measurement of human performance on Caltech 101 and Caltech 256 is not yet available).

Using datasets such as Caltech 101 and Caltech 256 has helped us make progress in a number of directions:

1. We have been able to make progress towards categorical recognition, as opposed to recognition of specific objects (my coffee mug, your face).
2. Algorithms are now tested for a variety of appearance statistics. Bonsai trees look different from baseball bats. Also, their pattern of variability is rather different. We may find out which approaches are specific to a category (e.g. faces) and which approaches generalize to multiple categories.
3. A number of different strategies have been proposed for extracting useful visual information from images. The so-called ‘features’. We have understood that visual features are shared across many categories.
4. Researchers have become aware that learning categories from a small number of training examples is a crucial problem [Fei-Fei et al., 2004].
5. Ideas have come about for making recognition fast.

While we should be happy about the good clip of progress of classification on datasets such as Caltech 256, we should be aware of limitations to the current approaches and challenge that lay ahead of us:

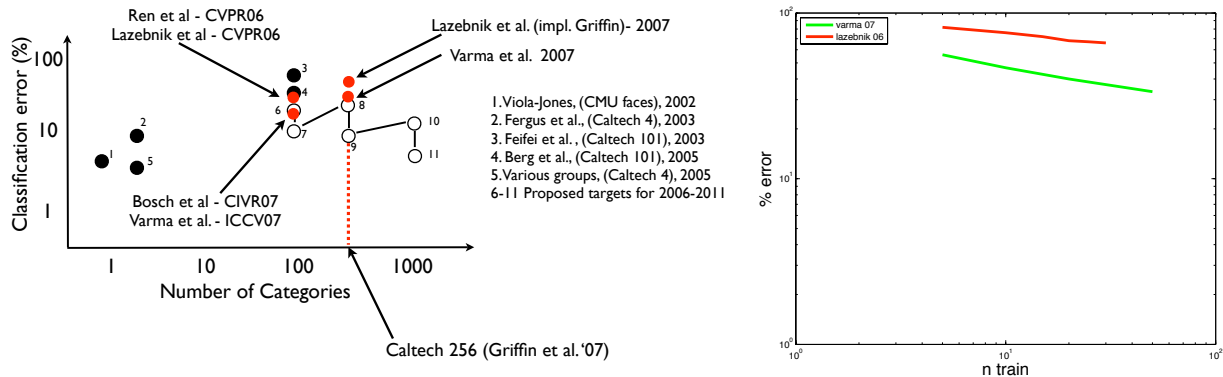


Figure 3: How well are we doing? (Left) Classification performance has seen steady improvement in the last few years, both in the number of categories on which algorithms are tested and in classification error rates. (Right) Performance of the best 2006 [Lazebnik et al., 2006] and the best 2007 algorithm [Varma, 2007] are compared here (classification error rates vs number of training examples). One may notice the significant year-on-year progress (see also left panel). Extrapolation enthusiasts may calculate that 10^8 training examples would be sufficient to achieve 1% error rates with current algorithms. Furthermore, if the pace of year-on-year progress is constant on this log scale chart, 1% error rates with 30 training examples will be achieved in 8-10 years.

1. We are still far from ‘good performance’. 1 % error rates on Caltech 256 are not around the corner.
2. This is still a ‘classification’ task, not a ‘naming’ task. Localization, as well as occlusion and clutter remain challenging problems.
3. There is no reason to believe that current approaches generalize across viewpoint: Once we learn ‘car’ from the side, our algorithms will not recognize ‘car’ when viewed frontally. Scale-invariance remains a challenge.
4. Current algorithms will not scale well with the number of categories. Their cost increases linearly at best.
5. Current algorithms do not scale well with the size in the image. Again: scaling is linear or super-linear at the moment.
6. There is yet no notion of the ‘distance’ between categories. We know that learning should be ‘generative’ for distant categories and ‘discriminative’ for similar categories citeholubWP05 but little work has been done on this problem.

6 Visual learning

Whether our goal is verification or full-fledged scene understanding, models of visual categories must be acquired. Unlike stereoscopy, motion and shape perception, where geometry and physics constrain the problem, there is no fundamental law telling us what a frog and a cell-phone should look like. A vast amount of information has to be learnt either from experts or from training examples.

The knowledge that may be acquired from training examples goes beyond phenomenological models of objects, materials and scenes. An overall organization, or ‘taxonomy’, of our visual knowledge is also desirable, as well as the statistics of co-occurrence of environments, objects and materials.

Depending on the context in which learning takes place, different learning strategies are called for. A great number of tantalizing problem face us:

1. One-shot learning. Learning new categories from a handful training examples would be very useful: training examples are difficult to come by. As Don Geman is fond of saying, “The interesting limit is $n \rightarrow 1$ ”. Bayesian [Fei-Fei et al., 2004, Fei-Fei et al., 2005] and other approaches [Holub et al., 2008] seem promising.
2. Incremental learning. While most efforts have understandably focussed on batch learning, it is clear that both animal and machine are better off if they can improve their knowledge of the world as new training examples come by. An additional benefit is that incremental approaches often offer faster learning algorithms.
3. Category formation. When do we have enough evidence to form a new category? How do we bootstrap knowledge about related categories in forming a new one?
4. Taxonomization. How do we organize our visual knowledge? Can we discover the relationship between different categories and group them into ‘broader’ categories? Can we take advantage of shared properties to make learning more effective, to speed up categorization and naming, to simplify the front-end of a visual system by sharing related mechanisms [Torralba et al., 2004]?
5. Human-machine interfaces. Humans are expert recognizers, but their time is precious. A machine ought to harvest information from human experts without wasting their time [Kapoor et al., 2007]. Could existing human-provided information (e.g. the Google image search engine) be harvested for useful training information [Barnard et al., 2003, Fergus et al., 2005]?
6. Unsupervised learning. Can machines learn from collections of complex images and movies (e.g. the entire archive of all TV soap operas) without any further help [Weber et al., 2000, Russell et al., 2006]?

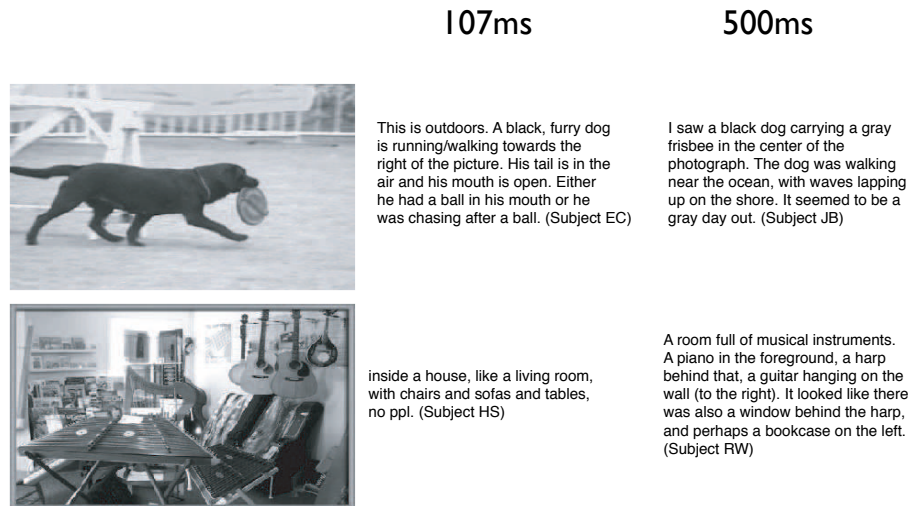


Figure 4: Human observers find it natural to provide a few sentences describing what they saw in a complex image. This is true even when the image is seen only briefly. (Adapted from [Fei-Fei et al., 2007])

7 Towards scene understanding

Is the task of ‘image understanding’ well defined? One might take a hard-nosed point of view and say that visual recognition is irrelevant unless it is directed at some goal (e.g. feeding, fighting or mating). In that case one could claim that there is no such thing as ‘scene understanding’; rather, there is preparation for an action. According to this line of thinking, visual recognition cannot be studied in the abstract, but it ought to be understood as part of a complete perception-action loop. There are indeed successful examples of visual recognition systems conceived with a specific task in mind and would not generalize beyond that, fingerprint recognition and face detection, for instance. However, for humans it is an easy and natural to give accurate verbal descriptions of images even when no task is specified and when images are shown for a brief moment (see Fig. 4). This suggests that a category-level description of the scene may be useful for carrying out a diverse set of unrelated tasks. Therefore, producing informative general-purpose high-level scene descriptions is a worthwhile academic goal. Solving this task will entail understanding the nature of these descriptions, as well as how to produce them.

What does a task-independent image description look like? We do not know. Language provides a degree of guidance, but it is likely that such a description might be more informative than what can be easily verbalized. A first step towards this is naming: producing the full list of ‘things’ we see in a scene, and where they are. That would include ‘global’ labels (such as ‘office’ and ‘kitchen’), as well as objects (such as ‘penguin’ and ‘inkwell’) and materials (such as ‘leather’ and ‘sand’). But naming is clearly just the beginning: we also



Figure 5: Different objects in an image have different importance for human observers, and this affects the probability of report. (Adapted from [Spain and Perona, 2007]). Photographs by S. Shore.

need to be able to compute and express properties, actions and relationships. I.e. we need not only names, but also verbs, adjectives and adverbs, and we need to combine all these together into meaningful descriptions.

Researchers working on recognition at the moment ignore the geometry of the environment. At some point both ‘geometry’ and ‘recognition’ must come together for proper scene understanding. Coming up with the simple statement “an apple is on the table” is will require both.

The relative importance of different statements that one can make about a scene is also an issue that needs to be understood (See Fig. 5). The description ‘sky above ground’ or ‘ceiling above floor’ might be the most appropriate for any scene, if we look at the relative size and visibility of different areas in the picture. These are, of course, hardly informative descriptions. How do we select a most informative subset of the exceedingly large number of statements one could make about a scene?

8 Conclusions

Visual recognition is one of the most exciting challenges of modern engineering. Approaching visual recognition makes us think hard about fundamental and diverse issues, such as the geometry and photometry of visual representations, the statistical properties of the world, unsupervised learning, categorization and model selection, efficient approximate optimization

and search, and the link between perception and cognition. Understanding the computational foundations of visual recognition helps us understand the brain. Machines endowed with visual recognition would enable unprecedented applications and change our lives.

We have made progress in the recent past: it is fair to say that now we have some understanding of the problem, as well as ideas of how to approach it. This was not the case just ten years ago. We are still far from addressing many of the ‘known unknowns’: naming, sublinear scaling with image size and number of categories, viewpoint and lighting invariance, and extrapolation.

There are also ‘unknown unknowns’: many believe that ‘scene understanding’ is the grand goal, but we do not yet know what form scene descriptions should take, nor how to approach producing such descriptions. Similarly, ‘taxonomies’ organizing our visual knowledge and relating similar categories are believed to be useful, but we do not yet have a clear view of the organizing principles for these taxonomies and of possible algorithms for producing them.

References

- [Barnard et al., 2003] Barnard, K., Duygulu, P., Forsyth, D. A., de Freitas, N., Blei, D. M., and Jordan, M. I. (2003). Matching words and pictures. *Journal of Machine Learning Research*, 3:1107–1135.
- [Fei-Fei et al., 2004] Fei-Fei, L., Fergus, R., and Perona, P. (2004). Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *IEEE CVPR Workshop of Generative Model Based Vision (WGMBV)*.
- [Fei-Fei et al., 2005] Fei-Fei, L., Fergus, R., and Perona, P. (2005). One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- [Fei-Fei et al., 2007] Fei-Fei, L., Iyer, A., Koch, C., and Perona, P. (2007). What do we perceive in a glance of a real-world scene? *Journal of Vision*, 7(1534-7362 (Electronic)):1–29.
- [Fergus et al., 2005] Fergus, R., Fei-Fei, L., Perona, P., and Zisserman, A. (2005). Learning object categories from google’s image search. In *ICCV*, pages 1816–1823.
- [Fleuret and Geman, 2001] Fleuret, F. and Geman, D. (2001). Coarse-to-fine face detection. *International Journal of Computer Vision (IJCV)*, 41(1-2):85–107.
- [Griffin et al., 2007] Griffin, G., Holub, A., and Perona, P. (2007). Caltech-256 object category dataset. Technical Report CNS-TR-2007-001, California Institute of Technology.
- [Holub et al., 2008] Holub, A., Welling, M., and Perona, P. (2008). Hybrid generative-discriminative visual categorization. *International Journal of Computer Vision*, 77(1-3):239–258.

- [Kapoor et al., 2007] Kapoor, A., Grauman, K., Urtasun, R., and Darrell, T. (2007). Active learning with gaussian processes for object categorization. In *International Conference on Computer Vision*. IEEE.
- [Lazebnik et al., 2006] Lazebnik, S., Schmid, C., and Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2169–2178.
- [Russell et al., 2006] Russell, B. C., Freeman, W. T., Efros, A. A., Sivic, J., and Zisserman, A. (2006). Using multiple segmentations to discover objects and their extent in image collections. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1605–1614.
- [Spain and Perona, 2007] Spain, M. and Perona, P. (2007). Measuring and predicting importance of objects in our visual world. Caltech CNS Technical Report 2007-002, California Institute of Technology.
- [Torralba et al., 2004] Torralba, A. B., Murphy, K. P., and Freeman, W. T. (2004). Sharing features: Efficient boosting procedures for multiclass object detection. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 762–769.
- [Varma, 2007] Varma, M. (2007). Entry in the caltech 256 competition. Visual Recognition Challenge Workshop.
- [Viola and Jones, 2004] Viola, P. and Jones, M. J. (2004). Robust real-time face detection. *Int. J. Comput. Vision*, 57(2):137–154.
- [Weber et al., 2000] Weber, M., Welling, M., and Perona, P. (2000). Towards automatic discovery of object categories. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, Hilton Head Island, USA.