BEWARE: These are preliminary notes. In the future, they will become part of a textbook on Visual Object Recognition.

Lecture 12: Computer Vision

If you can do it, a computer can do it too. Significant progress has been made over the last decade in teaching computers to perform multiple tasks that were traditionally thought to be the domain of humans. Any desktop computer can play chess competitively and the best computers can beat the world's chess champion. IBM's Watson has thrived in the trivia-like game of Jeopardy. And while imperfect, Siri and related systems are making enormous strides in becoming the world's best assistants.

In the domain of vision, computational algorithms are already able to perform certain tasks such as recognizing digits in a fully automatic fashion at human performance level and demonstrate reasonable performance in other tasks such as detecting faces for focusing on in digital cameras. In several other tasks, humans still outperform the most sophisticated current algorithms but the gap between machines and humans in vision tasks is closing rapidly. Here we provide an overview of several computer vision systems, particularly in the context of pattern recognition problems and describe what machine vision systems can and cannot do,

12.1 Applications

There is no shortage of applications where automatic or semi-automatic algorithms are being explored in computer vision. Here are a few examples:

(A) Intelligent content-based search. Searching for images in the web by content will open the doors to a large number of applications. Facebook users can already experiment with prototypes that let them search for people. One may be able to look for images that are similar to a search query in terms of content. Blind people may be able to point their phones and find out where they are and how to navigate.

(B) Prototype cars that can navigate automatically rely heavily on algorithms to detect pedestrians, other cars, other objects and road conditions.

(C) ATM machines may be able to recognize their customers. Cars and houses may recognize their owners.

(D) Security screening in places like airports may benefit from automatic recognition systems.

(E) Several clinical problems are based on pattern recognition and computers may soon help doctors to make informed decisions based on their understanding of patterns.

12.2 Computer vision tasks

Algorithms have been developed to address several interrelated problems in machine vision. While some of the boundaries are blurred in several applications, it is useful to think about the following tasks:

(A) Object detection. For example, a digital camera may require detecting the presence and location of a face in an image for focusing. Face detection may thrive without solving the problem of recognition.

(B) Object segmentation. In natural images, it may be of interest to separate an object from the background. For example, it may be important to detect the location of a tumor in an image. Or to detect the presence of a tank in a camouflaged image.

(C) Object recognition. Recognizing objects can often be thought of as associating the image with labels. These labels may refer to the identity of the object (e.g. given a face, who is it?) or the object's category (e.g. is there an animal in this image?).

(D) Object verification. In some cases, it may be of interest to evaluate whether two images are the same or not.

12.3 Object segmentation

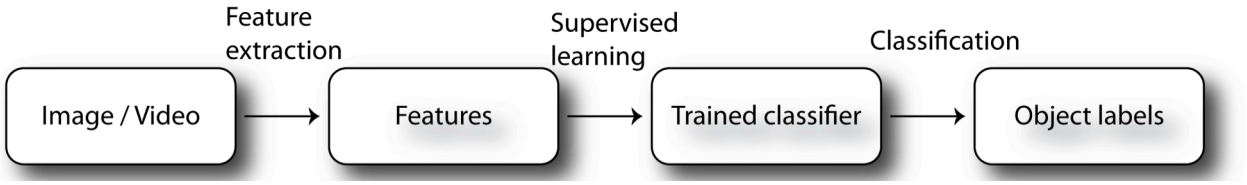
Given a natural scene, humans (and other species) are quite good at being able to characterize and localize different objects embedded in complex backgrounds. The fact that this is not a trivial problem is highlighted by the ubiquitous use of camouflage in the animal world. Particularly for objects that are not moving, matching colors, contrast and textures can help animals avoid predators or at least buy sufficient time for escape. Basic aspects of segmentation may depend on adequately detecting edges. However, more complex problems often involve a deeper understanding of the interrelationships among different object parts. A typical case involves recognizing a zebra as a whole animal as opposed to thinking of each stripe as a separate object.

Some algorithms require recognition prior to segmentation while other algorithms use segmentation to guide recognition in complex scenes. To avoid this chicken-and-egg dilemma, it is tempting to speculate that certain aspects of bottom-up recognition and segmentation could occur (or at least) start independently of each other, using overlapping neuronal circuits. Top-down signals may then combine segmentation and recognition in synergistic fashion. For examples of object segmentation algorithms see (Borenstein et al., 2004).

12.4 A general scheme for object recognition

Figure 12.1 illustrates a typical approach in computer vision efforts. Consider a series of N labeled images $(\boldsymbol{x}_i, y_i)$ where i=1,…,N, $\boldsymbol{x}$ is a matrix representing the image and y is a label (e.g. face present or not). A set of features f is extracted from the images: $\boldsymbol{f}_i=g(\boldsymbol{x}_i)$. Those features may include properties such as edges, principal components, etc. How those features are chosen is one of the key aspects that differentiates computer vision algorithms. A supervised learning scheme is then used to learn the map between those

Figure 12.1. A general scheme for object recognition. Features are extracted from an image (or video). Those features are used to train a classifier via supervised learning. The resulting classification boundary is used with novel images (different from the ones used during training) to assign object labels to images.

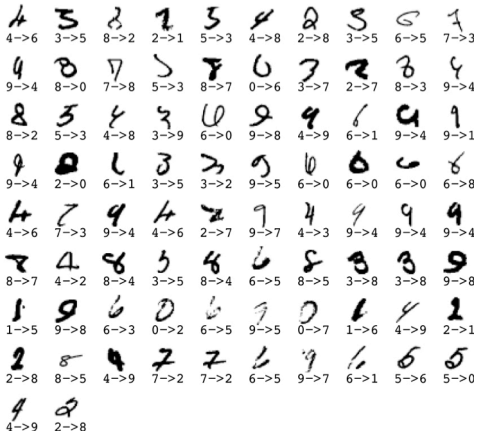


features and labels (Poggio and Smale, 2003; Meyers and Kreiman, 2011; Singer and Kreiman, 2011). For example, a support vector machine (SVM) classifier with a linear kernel may be used to learn the structure of the data and labels. A cross-validation procedure is followed by separating the data into a training set and a test set to avoid overfitting. After training, the algorithm is evaluated with the images in the test set. By using different algorithms applied to the same data, the merits of alternative approaches can be quantitatively compared.

12.5 A successful example: digit recognition

Recognizing hand-written digits constitutes an example where computers have reached high accuracy, almost comparable to human levels (e.g. (LeCun et al., 1998)). Figure 12.2 shows an example of the errors made by an early attempt at recognizing hand-written digits. The overall error rate of this algorithm was <2%. Several of those errors are not trivial to recognize and humans could make mistakes as well.

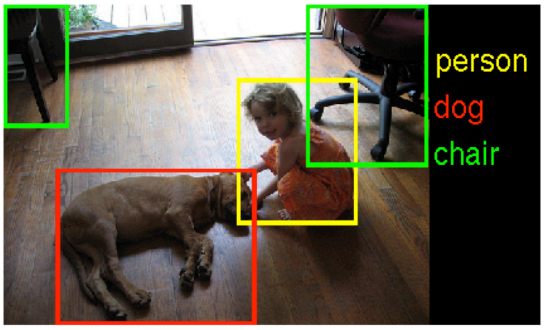
Figure 12.2. Example of digit recognition mistakes by the algorithm in LeCun et al 1988. Below each digit, the image shows the true label and the computer label.



12.6 Image recognition competitions

There are several computer vision competitions with large data sets consisting of labeled images. One such competition is called the Imagenet Large Scale Visual Recognition Challenge (ILSVRC) (Russakovsky et al., 2014). The 2014 instantiation of the object classification part of this challenge included 1000 object classes, 1,281,413 images for training (732-1300 images per class) and 100,000 images for testing (100

Figure 12.3. Example labeled image from the validation set in the 2013 ILSVRC competition. [Image source: http://www.image-net.org/challenges/LSVRC/2013/]



images per class). This competition also includes other tasks beyond classification including object detection and localization. To give an idea of performance, the winning team in the object classification part of the challenge achieved an error rate slightly above 6%. This is quite impressive considering that there were 1000 classes. It should be noted, though, that the results of these competitions are often reported in a somewhat strange way by allowing the models 5 changes to get it right and reporting the results as correct if any of those 5 predictions are correct. This makes it a bit more difficult to directly compare against human performance (Borji and Itti, 2014). Another aspect of machine vision that has also been highlighted is the difficulty in interpreting how the machine classifies objects and investigators have reported puzzling examples where minimal changes to an image drastically change the predicted class (Szegedy et al., 2014). With that said, the results are still quite remarkable and they show rapid progress in teaching machines to recognize objects.

References

Borenstein E, Sharon E, Ullman S (2004) Combining Top-Down and Bottom-Up Segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Washington, DC.

Borji A, Itti L (2014) Human vs. computer in scene and object recognition. In: CVPR.

LeCun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. Proc of the IEEE 86:2278-2324.

Meyers EM, Kreiman G (2011) Tutorial on Pattern Classification in Cell Recordings. In: Understanding visual population codes (Kriegeskorte N, Kreiman G, eds). Boston: MIT Press.

Poggio T, Smale S (2003) The mathematics of learning: dealing with data. Notices of the AMS 50:537-544.

Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang S, Karpathy A, Khosla A, Bernstein M, Berg A, Fei-Fei L (2014) ImageNet Large Scale Visual Recognition Challenge. In: CVPR: arXiv:1409.0575, 2014.

Singer J, Kreiman G (2011) Introduction to Statistical Learning and Pattern Classification. In: Understanding visual population codes (Kriegeskorte N, Kreiman G, eds). Boston: MIT Press.

Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, Fergus R (2014) Intriguing properties of neural networks. In: International Conference on Learning Representations.