The Emergence of Visual Categories - A Computational Perspective

Pietro Perona[1]

California Institute of Technology
Pasadena, CA 91001
perona@caltech.edu,

Abstract. When we are born we do not know about sailing boats, frogs, cell-phones and wheelbarrows. By the time we reach school age we can easily recognize these categories of objects and many more using our visual system; by some estimates, we learn around 10 new categories per day with minimal supervision during the first few years of our lives. How can this happen? I will outline a computational approach to the problem of representing the visual properties of object categories, and of learning such models without supervision from cluttered images. Both static images of objects and dynamic displays such as the ones generated by human activity are handled by the theory. Its properties will be exemplified with experiments on a variety of categories.

My collaborators and I are interested in representation, detection/recognition and learning of visual object categories. We are developing a probabilistic approach for addressing these issues in a principled and consistent manner. I will briefly review our work to date giving pointers to the relevant publications.

Representation – Representatives of a category (e.g. cars) look similar enough to be grouped together, yet they may be quite different in shape, presence or absence of features or distinctive parts (e.g. glasses and mustache on human faces). Even when present these features may look quite different from object to object. The challenge of representation is capturing what is similar between objects belonging to the same category, while allowing for the necessary variability. To this end, we use probabilistic models. Each object is represented as a constellation of parts with distinctive appearance (e.g. eyes, nose, mouth) which appear in a given mutual position. The model of a class therefore includes a number of ingredients: (a) the appearance of the features associated with each part may be represented either with an ensemble of jets [1, 2], a fixed template [3, 4] or with a mixture model [5]; (b) this appearance allows one to synthesize appropriate detectors, each one of which is characterized by the probability that, given the presence of an object, the corresponding part will be detected (the part may of course not be detected due to failure of the detector, occlusion or simply absence of the feature – all these causes are described by a single parameter); (c) The mutual position of the parts is described by an appropriate probability density

J. Bigun and T. Gustavsson (Eds.): SCIA 2003, LNCS 2749, pp. 1–3, 2003.

function which is invariant to translation, rotation and scale [6, 3], or to affine deformations [7]. The statistics of detection false alarms arising from background clutter should be modeled as well [3, 4]. We may represent dynamic displays arising from humans performing actions (walking, reaching for an object...) in a similar fashion: as constellations of features in motion [8, 9].

So far we have developed only 2D view-based models. In order to deal with the multiplicity of appearances that an object generates under different viewing directions therefore we choose to glue together multiple 2D views. Our limited experiments (only faces and people walking!) show that our models are reasonably invariant with respect to viewpoint, suggesting that a coarse sampling of the viewing sphere may be sufficient in practice [10, 8].

Detection / Recognition – We approach detection and recognition as statistical decision problems: given the observed data what is the likelihood that a given object is present, and how does this compare with the likelihood that the same object is not in the scene? The exact expression for these decisions may be computed explicitly from our probabilistic model [6, 11]. If one is willing to make a few mild approximations, it is fast to compute and optimal from a probabilistic standpoint. Objects may thus be detected quickly in cluttered scenes while treating the variability of object appearance, clutter, occlusion, scale, translation and rotation in a principled manner.

Learning – The most challenging problem is learning. We wish to develop machine vision algorithms that are able to learn new object categories with minimal human supervision. The current state of the art in machine vision is unfortunately very far from this goal: most if not all existing algorithms require a human to patiently segment away spurious clutter and align the features (e.g. eyes, nose and mouth) of the training examples by scaling, translation, rotation and possibly affine deformations. We have proposed a maximum-likelihood algorithm that successfully trains probabilistic models of object categories without supervision [4] (i.e. in the absence of segmentation from clutter and in the absence of alignment). The current version is scale-invariant as well [5]. In the earlier version [4] the appearance of the features is learned first, and the shape of the model is learned in a second step. In the current version [5] both appearance and shape are learned at the same time. Our current experiments involve learning one object category at a time. It is desirable to generalize our approach to learning multiple categories at once - in one of our publications we describe an approach and test it in a limited setting [12]. Our approach to learning human motions is described in [13, 14].

References

1. Leung, T., Burl, M., Perona, P.: Finding faces in cluttered scenes using random labelled graph matching. In: Proc. 5^{th} Int. Conf. on Computer Vision, Cambridge, Mass (1995) 637–644

2. Burl, M., Leung, T., Perona, P.: Face localization via shape statistics. In: Proc. Intl. Workshop on automatic face and gesture recognition, Zurich, IEEE Computer Soc. (1995) 154–159
3. Burl, M., Weber, M., Perona, P.: A probabilistic approach to object recognition using local photometry and global geometry. In: Proc. 5^{th} Europ. Conf. Comput. Vision, Burkhardt and Neumann (Ed.), LNCS-Series Vol. 1406-1407, Springer-Verlag. (1998)
4. M. Weber, M.W., Perona, P.: Unsupervised learning of models for recognition. (In: Proc. 6th Europ. Conf. Comp. Vis., ECCV2000, Dublin, Ireland, June 2000)
5. R. Fergus, P.P., Zisserman, A.: Object class recognition by unsupervised scale-invariant learning. (In: IEEE Comp. Society Conf. on Computer Vision and Pattern Recognition, CVPR 03, Madison, WI, June 2003)
6. Burl, M., Perona, P.: Recognition of planar object classes. In: Proc. IEEE Comput. Soc. Conf. Comput. Vision and Pattern Recogn., San Francisco (1996)
7. Leung, T., Burl, M., Perona, P.: Probabilistic affine invariance for recognition. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition, Santa Barbara, CA (1998) 678–684
8. Song, Y., Goncalves, L., Bernardo, E.D., Perona, P.: Monocular perception of biological motion - detection and labeling. In: Proceedings of International Conference on Computer Vision. (1999) 805–812
9. Song, Y., Goncalves, L., Perona, P.: Monocular perception of biological motion - clutter and partial occlusion. In: Proc. ECCV. Volume 2. (2000) 719–733
10. M. Weber, W. Einhaeuser, M.W., Perona, P.: Viewpoint-invariant learning and detection of human heads. (In: Proc. 4th Int. Conf. Autom. Face and Gesture Rec., FG2000, Grenoble, France, March 2000)
11. Burl, M., Perona, P.: Using hierarchical shape models to spot keywords in cursive handwriting. (In: IEEE Comp. Society Conf. on Computer Vision and Pattern Recognition, CVPR 98, Santa Barbara, CA, June 1998)
12. M. Weber, M.W., Perona, P.: Towards automatic discovery of object categories. (In: Proc. IEEE Comp. Soc. Conf. Comp. Vis. and Pat. Rec., CVPR2000, June 2000)
13. Y. Song, L.G., Perona, P.: Unsupervised learning of human motion models. (In: Proc. of NIPS 2001) 1287–1294
14. Y. Song, L.G., Perona, P.: Unsupervised learning of human motion. IEEE Trans. Pattern Anal. Mach. Intell. **25** (2003)