Object Recognition

Ming-Hsuan Yang
University of California at Merced
http://faculty.ucmerced.edu/mhyang

SYNONYMS

Object Identification, Object Labeling.

DEFINITION

Object recognition is concerned with determining the identity of an object being observed in the image from a set of known labels. Oftentimes, it is assumed that the object being observed has been detected or there is a single object in the image.

HISTORICAL BACKGROUND

As the holy grail of computer vision research is to tell a story from a single image or a sequence of images, object recognition has been studied for more than four decades [9] [22]. Significant efforts have been paid to develop representation schemes and algorithms aiming at recognizing generic objects in images taken under different imaging conditions (e.g., viewpoint, illumination, and occlusion). Within a limited scope of distinct objects, such as handwritten digits, fingerprints, faces, and road signs, substantial success has been achieved. Object recognition is also related to content-based image retrieval and multimedia indexing as a number of generic objects can be recognized. In addition, significant progress towards object categorization from images has been made in the recent years [17]. Note that object recognition has also been studied extensively in psychology, computational neuroscience and cognitive science [9, 4].

SCIENTIFIC FUNDAMENTALS

Object recognition is one of the most fascinating abilities that humans easily possess since childhood. With a simple glance of an object, humans are able to tell its identity or category despite of the appearance variation due to change in pose, illumination, texture, deformation, and under occlusion. Furthermore, humans can easily generalize from observing a set of objects to recognizing objects that have never been seen before. For example, kids are able to generalize the concept of “chair” or “cup” after seeing just a few examples. Nevertheless, it is a daunting task to develop vision systems that match the cognitive capabilities of human beings, or systems that are able to tell the specific identity of an object being observed. The main reasons can be attributed to the following factors: relative pose of an object to a camera, lighting variation, and difficulty in generalizing across objects from a set of exemplar images. Central to object recognition systems are how the regularities of images, taken under different lighting and pose conditions, are extracted and recognized. In other words, all these algorithms adopt certain representations or models to capture these characteristics, thereby facilitating procedures to tell their identities. In addition, the representations can be either 2D or 3D geometric models. The recognition process, either generative or discriminative, is then carried out by matching the test image against the stored object representations or models.

Geometry-based approaches

Early attempts on object recognition were focused on using geometric models of objects to account for their appearance variation due to viewpoint and illumination change. The main idea is that the geometric description of a 3D object allows the projected shape to be accurately predicated in a 2D image under projective projection, thereby facilitating recognition process using edge or boundary information (which is invariant to certain illumination change). Much attention was made to extract geometric primitives (e.g., lines, circles, etc.) that are invariant to viewpoint change [13]. Nevertheless, it has been shown that such primitives can only be reliably extracted under limited conditions (controlled variation in lighting and viewpoint with certain occlusion). An excellent review on geometry-based object recognition research by Mundy can also be found in [12].

Appearance-based algorithms

In contrast to early efforts on geometry-based object recognition works, most recent efforts have been centered on appearance-based techniques as advanced feature descriptors and pattern recognition algorithms are developed [8]. Most notably, the eigenface methods have attracted much attention as it is one of the first face recognition systems that are computationally efficient and relatively accurate [21]. The underlying idea of this approach is to compute eigenvectors from a set of vectors where each one represents one face image as a raster scan vector of gray-scale pixel values. Each eigenvector, dubbed as an eigenface, captures certain variance among all the vectors, and a small set of eigenvectors captures almost all the appearance variation of face images in the training set. Given a test image represented as a vector of gray-scale pixel values, its identity is determined by finding the nearest neighbor of this vector after being projected onto a subspace spanned by a set of eigenvectors. In other words, each face image can be represented by a linear combination of eigenfaces with minimum error (often in the L2 sense), and this linear combination constitutes a compact reorientation. The eigenface approach has been adopted in recognizing generic objects across different viewpoints [14] and modeling illumination variation [2].

As the goal of object recognition is to tell one object from the others, discriminative classifiers have been used to exploit the class specific information. Classifiers such as k-nearest neighbor, neural networks with radial basis function (RBF), dynamic link architecture, Fisher linear discriminant, support vector machines (SVM), sparse network of Winnows (SNoW), and boosting algorithms have been applied to recognize 3D objects from 2D images [16] [6] [1] [18] [19]. While appearance-based methods have shown promising results in object recognition under viewpoint and illumination change, they are less effective in handling occlusion. In addition, a large set of exemplars needs to be segmented from images for generative or discriminative methods to learn the appearance characteristics. These problems are partially addressed with parts-based representation schemes.

Feature-based algorithms

The central idea of feature-based object recognition algorithms lies in finding interest points, often occurred at intensity discontinuity, that are invariant to change due to scale, illumination and affine transformation (a brief review on interest point operators can be found in [8]). The scale-invariant feature transform (SIFT) descriptor, proposed by Lowe, is arguably one of the most widely used feature representation schemes for vision applications [8]. The SIFT approach uses extrema in scale space for automatic scale selection with a pyramid of difference of Gaussian filters, and keypoints with low contrast or poorly localized on an edge are removed. Next, a consistent orientation is assigned to each keypoint and its magnitude is computed based on the local image gradient histogram, thereby achieving invariance to image rotation. At each keypoint descriptor, the contribution of local image gradients are sampled and weighted by a Gaussian, and then represented by orientation histograms. For example, the 16×16 sample image region and 4×4 array of histograms with 8 orientation bins are often used, thereby providing a 128-dimensional feature vector for each keypoint. Objects can be indexed and recognized using the histograms of keypoints in images. Numerous applications have been developed using the SIFT descriptors, including object retrieval [20] [15], and object category discovery [5].

Although the SIFT approach is able to extract features that are insensitive to certain scale and illumination change, vision applications with large base line change entail the need of affine invariant point and region operators [11]. A performance evaluation among various local descriptors can be found in [10], and a study on affine region detectors is presented in [11]. Finally, SIFT-based methods are expected to perform better for objects with rich texture information as sufficient number of keypoints can be extracted. On the other hand, they also require sophisticated indexing and matching algorithms for effective object recognition [8] [17].

KEY APPLICATIONS

Biometric recognition, and optical character/digit/document recognition are arguably the most widely used applications. In particular, face recognition has been studied extensively [23] for decades and with large scale ongoing efforts. On the other hand, biometric recognition systems based on iris or fingerprint as well as as handwritten digit have become reliable technologies [7] [3]. Other object recognition applications include surveillance, industrial inspection, content-based image retrieval (CBIR), robotics, medical imaging, human computer interaction, and intelligent vehicle systems, to name a few.

FUTURE DIRECTIONS

With more reliable representation schemes and recognition algorithms being developed, tremendous progress has been made in the last decade towards recognizing objects under variation in viewpoint, illumination and under partial occlusion. Nevertheless, most working object recognition systems are still sensitive to large variation in illumination and heavy occlusion. In addition, most existing methods are developed to deal with rigid objects with limited intra-class variation. Future research will continue searching for robust representation schemes and recognition algorithms for recognizing generic objects.

DATA SETS

Numerous face image sets are available on the web

FERET face data set:
`http://www.itl.nist.gov/iad/humanid/feret/`

- UMIST data set:
`http://images.ee.umist.ac.uk/danny/database.html`
- Yale data set:
`http://cvc.yale.edu/projects/yalefacesB/yalefacesB.html`
- AR data set:
`http://cobweb.ecn.purdue.edu/%7Ealeix/aleix_face_DB.html`
- CMU PIE data set:
`http://www.ri.cmu.edu/projects/project_418.html`

There are several large data sets for object recognition experiments,

COIL data set:
`http://www1.cs.columbia.edu/CAVE/software/softlib/coil-100.php`

- CalTech data sets:
`http://www.vision.caltech.edu/html-files/archive.html`
- PASCAL visual object classes:
`http://www.pascal-network.org/challenges/VOC/`

URL TO CODE

There are a few excellent short courses on object recognition in recent conferences available on the web.

"Recognition and matching based on local invariant features" by Schmid and Lowe in IEEE Conference on Computer Vision and Pattern Recognition 2003:
`http://lear.inrialpes.fr/people/schmid/cvpr-tutorial03/`

- "Learning and recognizing object categories" by Fei-Fei, Fergus and Torralba in IEEE International Conference on Computer Vision 2005:
  http://people.csail.mit.edu/torralba/shortCourseRLOC/
- "Recognizing and Learning Object Categories: Year 2007" by Fei-Fei, Fergus and Torralba in IEEE Conference on Computer Vision and Pattern Recognition 2005:
  http://people.csail.mit.edu/torralba/shortCourseRLOC/

Sample code for face recognition and SIFT descriptors:

Face recognition:
http://www.face-rec.org/

- Lowe's sample SIFT code:
  http://www.cs.ubc.ca/~lowe/keypoints/.
- MATLAB implementation of SIFT descriptors by Vedaldi:
  http://vision.ucla.edu/~vedaldi/code/sift/sift.html
- libsift by Nowozin:
  http://user.cs.tu-berlin.de/~nowozin/libsift/

Grand challenge in object recognition:

NIST face recognition grand challenge:
http://www.frvt.org/FRGC/

- NIST multiple biometric grand challenge:
  http://face.nist.gov/mbgc/
- PASCAL visual object classes challenge 2007:
  http://www.pascal-network.org/challenges/VOC/voc2007/index.html

CROSS REFERENCE

Object detection.

RECOMMENDED READING

[1] P. Belhumeur, J. Hespanha, and D. Kriegman. Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):711–720, 1997.

[2] P. Belhumeur and D. Kriegman. What is the set of images of an object under all possible illumination conditions. *International Journal of Computer Vision*, 28(3):1–16, 1998.

[3] J. Daugman. Probing the uniqueness and randomness of iriscodes: Results from 200 billion iris pair comparisons. *Proceedings of the IEEE*, 94(11):1927–1935, 2006.

[4] S. Edelman. *Representation and recognition in vision.* MIT Press, 1999.

[5] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 264–271, 2003.

[6] M. Lades, J. C. Vorbrüggen, J. Buhmann, J. Lange, C. von der Malsburg, R. P. Würtz, and W. Konen. Distortion invariant object recognition in the dynamic link architecture. *IEEE Transactions on Computers*, 42:300–311, 1993.

[7] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

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

[9] D. Marr. *Vision.* W. H. Freeman and Company, 1982.

[10] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(10):1615–1630, 2005.

[11] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool. A comparison of affine region detectors. *International Journal of Computer Vision*, 65(1/2):43–72, 2006.

[12] J. Mundy. Object recognition in the geometric era: a retrospective. In J. Ponce, M. Hebert, C. Schmid, and A. Zisserman, editors, *Toward category-level object recognition*, pages 3–29. Springer-Verlag, 2006.

[13] J. Mundy and A. Zisserman, editors. *Geometric invariance in computer vision.* MIT Press, 1992.

[14] H. Murase and S. K. Nayar. Visual learning and recognition of 3-D objects from appearance. *International Journal of Computer Vision*, 14:5–24, 1995.

[15] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 2161–2168, 2006.

[16] T. Poggio and S. Edelman. A network that learns to recognize 3D objects. *Nature*, 343:263–266, 1990.

[17] J. Ponce, M. Hebert, C. Schmid, and A. Zisserman, editors. *Toward category-level object recognition.* Springer-Verlag, 2006.

[18] M. Pontil and A. Verri. Support vector machines for 3d object recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(6):637–646, 1998.

[19] D. Roth, M.-H. Yang, and N. Ahuja. Learning to recognize objects. *Neural Computation*, 14(5):1071–1104, 2002.

[20] J. Sivic and A. Zisserman. Video Google: a text retrieval approach to object matching in videos. In *Proceedings of IEEE International Conference on Computer Vision*, pages 1470–1477, 2003.

[21] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.

[22] S. Ullman. *High-level vision: Object recognition and visual recognition.* MIT Press, 1996.

[23] W. Zhao, R. Chellappa, A. Rosenfeld, and J. P. Phillips. Face recognition: A literature survey. *ACM Computing Surveys*, 35(4):399–458, 2003.