Workshop on Generic Object Recognition and Categorization

version October 7, 2004

Sven Dickinson
University of Toronto
Canada
sven@cs.toronto.edu

Aleš Leonardis
University of Ljubljana
Slovenia
alesl@fri.uni-lj.si

Bernt Schiele
Darmstadt University of Technology
Germany
schiele@informatik.tu-darmstadt.de

Contents

1. Overview

On June 27, 2004 the third "Workshop on Generic Object Recognition and Categorization" has been held in Washington D.C. with the generous support from the ECVision network. The workshop was held as one of the workshops of CVPR 2004 (IEEE Conference on Computer Vision and Pattern Recognition). Like the 1st and the 2nd such workshop, held in 1997 and 1999, respectively, the format of the workshop consisted of a set of invited talks and a panel discussion by leading object recognition researchers who have addressed the problem of object categorization.

Rather than having a mini-conference where people present their recent results we encouraged the invited speakers to present a retrospective of their past research, highlight their current research, as well as to share their vision for future research. Some of the questions we asked the invited speakers prior to the workshop were the following: "We'd like to know to what extent the object representations you've worked with scale up to categories? Which representations suit the task of generic object modeling? How much attention should be devoted to perceptual grouping? How should we address the problems of indexing and matching? What is the role of learning in solving these problems? What do you feel are the major challenges facing generic object recognition and categorization?"

The declared goal of the workshop was to take a broad view on the problem, including speakers from both computer vision and human vision areas. Furthermore, while traditional work in generic object recognition assumed higher level object descriptions (whether 2-D viewer-centered or 3-D object-centered), recent work closer to the appearance level has suggested that some lower level features may be sufficient for at least some categorization problems. On this representational spectrum, we also wanted a balance, including speakers from different camps. We explicitly stated that we do not want to promote a particular methodology, but to solicit feedback from the broadest collection possible of those working in generic object recognition and categorization, including those from the human vision community, who can bring psychophysical and neuroscience arguments to the table.

In the following we briefly motivate the problem of the workshop and then present the final program of the workshop, the speakers and the titles and abstracts of their pre-

sentations. We were very pleased that we were able to convince a set of higly distinctive speakers. As an indication of the success of the workshop the number of participants exceeded the expectations of the CVPR-organziers. In fact we estimate that the number of participants during the workshop day was up to 120 people[1]

2. Topic and Motivation

The capacity to categorize objects plays a crucial role for a cognitive and autonomous visual system in order to compartmentalize the huge numbers of objects it has to handle into manageable categories. Quite interestingly, for humans it was shown that entry-level categorization (i.e. Is this a dog/cat?) is much faster than recognition or identification (Is this my dog/cat?). These findings suggest that humans do a sort of coarse to fine categorization and recognition of objects.

Even though generic object recognition and classification have been one of the goals of computer vision since its beginnings, we are still far from achieving this goal. On the other hand, the identification of known objects in different poses and under novel viewing conditions has made significant progress recently. At the same time, impressive results have been achieved for the detection of canonical views of individual categories, such as faces, cars, pedestrians, and horses. While the more general task of multi-class object categorization is still unsolved, we have seen at recent conferences such as CVPR 2003 and ICCV 2003 that research in the area regains momentum and new approaches emerge.

Generic object recognition endeavors to recognize objects based on their coarse, prototypical shape. Although a popular topic in the 1970's, generic object recognition has given up the recognition spotlight over the years to such schemes as alignment, geometric invariant-based indexing, and more recently, appearance-based and local feature-based recognition. While all of these approaches have their advantages and disadvantages it is not clear what the role of different visual cues (such as contour, shape, color, texture, etc.) is, and what the role of object models are for generic object recognition. Traditionally, contour-, shape-, and part-based methods are considered most adequate for handling the generalization requirements needed for categorization tasks, even though most current object recognition and detection systems are appearance-based.

2.1. Organization and Workshop Format

In order to achieve the most stimulating discussions around the theme of generic object recognition and visual object categorization, 11 well known-researchers in the field with a record in the area of generic object recognition and visual object categorization have been invited. The workshop day has been concluded by a general discussion by all workshop participants about current and future trends in the field.

[1] as a matter of fact we had to change our assigned room twice – once just before starting the workshop and then again during the day due to the large number of partcipants interested in the workshop

3. Program

Table 1 contains the final program of the workshop. The following sections contain the titles and abstracts of the invited talks in the order of the program. Section **??** contains a summary of the panel discussion. The actual slides of most presentations can be downloaded from the following webpage: http://www.vision.ethz.ch/cvpr04-gorc.

Table 1. Final program of the workshop

8:50 – 9:00	**Introduction by the organizers**
9:00 – 10:30	**Calibration Session: Where are We Today?**
	Generic Shape Learning and Recognition Gérard Medioni, University of Southern California
	Object Recognition and Categorization: Some Lessons from Psychophysics, Neurobiology and Computer Vision Shimon Edelman, Cornell University
	Human Object Recognition – Do We Know More than We Did 20 Years Ago? Michael Tarr, Brown University
11:00 – 12:30	**Session: Neuroscience**
	Dimensions of Neural Shape Space C.E. Connor, John Hopkins Univeristy
	Computational Mechanisms of Generic Object Recognition and Categorization in Cortex Maximilian Riesenhuber, Georgetown University
	Representation of Object Images in the Monkey Inferotemporal Cortex Manabu Tanifuji, Riken Brain Science Institute, Japan
13:30 – 14:30	**Session: Image- and Learning-Based**
	End-to-End Learning of Object Categorization with Invariance Pose, Illumination, and Clutter Yan LeCun, New York University
	Recognizing Objects Using Shape Jitendra Malik, U. C. Berkeley
15:00 – 16:30	**Session: Shape-Based and Beyond**
	Toward True 3D Object Recognition Jean Ponce, Beckman Institute
	A Shockgraph Dissimilarity Metric for Object Recognition Benjamin Kimia, Brown University
	Highlevel Vision and Links to Language David Forsyth, U. C. Berkeley
16:30 – 18:00	**Panel**
	Jitendra Malik, U. C. Berkeley Gérard Medioni, University of Southern California Maximilian Riesenhuber, Georgetown University Michael Tarr, Brown University

3.1. Generic Shape Learning and Recognition

Gérard G. Medioni
University of Southern California
USA

Abstract We discuss the issues and challenges of generic object recognition. We argue that high-level, volumetric part-based descriptions are essential in the process of recognizing objects that might never have been observed before, and for which no exact geometric model is available. We discuss the representation scheme and its relationships to the three main tasks to solve: extracting descriptions from real images, under a wide variety of viewing conditions; learning new objects by storing their description in a database; recognizing objects by matching their description to that of similar previously observed objects.

References

- http://iris.usc.edu/home/iris/medioni/User.html
- Zerroug, M., and Medioni, G., The Challenge of Generic Object Recognition, Object Representation in Computer Vision'95, (pp. 217-232).
- Zerroug, M., and Nevatia, R., Volumetric Descriptions from a Single Intensity Image, IJCV (20), No. 1/2, 1996, pp. 11-42.
- Medioni, G.G., Franois, A.R.J., 3D Structures for Generic Object Recognition, ICPR'00 (Vol I: pp 30-37)
- Francois, A.R.J., Medioni, G.G., Generic Shape Learning and Recognition, Object Representation in Computer Vision'96 (p 287)
- Havaldar, P., Medioni, G.G., and Stein, F., Perceptual Grouping for Generic Recognition, IJCV (20), No. 1/2, 1996, pp. 59-80.

3.2. Object Recognition and Categorization: Some Lessons from Psychophysics, Neurobiology and Computer Vision

Shimon Edelman
Department of Psychology
Cornell University
Ithaca, NY 14853
USA

Abstract Much useful information about a visual object can be obtained by computing its similarities to a small number of reference shapes or prototypes, which, in turn, can be represented by their view spaces, interpolated from a handful of exemplar views. Such low-dimensional, hence computationally tractable, view-based representations support both the recognition of familiar shapes and the categorization of novel ones [1]. Apart from categorization, they can also be used in a variety of other tasks involving novel objects: viewpoint-insensitive recognition, recovery of a canonical view, and estimation of pose or of arbitrary novel views [2]. Predictions generated by this computational model concerning the cortical physiology of object representation in primates have been borne out by experiments (e.g., [3,4,5]). Moreover, its limitations vis-a-vis dealing with progressive shape change and with image translation (as well as other stimulus manipulations) resemble those of human subjects [6,7]. However, the absolute level of performance of the implemented system that had been based on this approach [8] fell short of the human standard. In this talk, I shall discuss possible approaches to closing this performance gap while keeping the model computationally feasible and biologically relevant.

References

1. S. Edelman. Representation is representation of similarity. Behavioral and Brain Sciences, 21:449-498, 1998.
2. S. Edelman and S. Duvdevani-Bar. Similarity-based viewspace interpolation and the categorization of 3D objects. In Proc. Similarity and Categorization Workshop, pages 75-81, Dept. of AI, University of Edinburgh, 1997.
3. N.K. Logothetis, J. Pauls, and T. Poggio. Shape recognition in the inferior temporal cortex of monkeys. Current Biology, 5:552-563, 1995.
4. D.J. Freedman, M. Riesenhuber, T. Poggio, and E.K. Miller. Categorical representation of visual stimuli in the primate prefrontal cortex. Science, 291:312-316, 2001.

5. H.Op de Beeck, J. Wagemans, and R. Vogels. Inferotemporal neurons represent low-dimensional configurations of parameterized shapes. Nature Neuroscience, 4:1244-1252, 2001.

6. S. Edelman. Representation and recognition in vision. MIT Press, Cambridge, MA, 1999.

7. M. Dill and S. Edelman. Imperfect invariance to object translation in the discrimination of complex shapes. Perception, 30:707-724, 2001.

8. S. Duvdevani-Bar and S. Edelman. Visual recognition and categorization on the basis of similarities to multiple class prototypes. Intl. J. Computer Vision, 33:201-228, 1999.

3.3. Human Object Recognition – Do We Know More than We did 20 Years Ago?

Michael J. Tarr
Department of Cognitive and Linguistic Sciences
Brown University
USA

Abstract The intensive study of the mechanisms of human object recognition arguably began 20 years ago, sparked, in part, by the publication of David Marrs landmark book, Vision. Since that time there has been an incredible number of behavioral and, more recently, neuroimaging, studies focusing on questions such as invariance and domain specificity. There has been one popular theory, two raucous debates, and at least three generations of new researchers entering the field. Yet for all this activity, we still lack a plausible (and detailed) model of generic object recognition that can explain even a fraction of the psychophysical and neural data we have collected. What is going on here? First, it is a hard problem. Second, some of the questions asked over the past two decades have probably been the wrong ones. For example, a great deal of energy was expended on whether human object recognition was viewpoint-invariant or viewpoint-dependent. The winner seems to beit depends. That is, there are cases where observers are able to identify objects invariantly across changes in viewpoint and there are cases where observers show dramatic dependency on viewpoint. Dismissing either as an exception outside the bounds of theory is a mistake: humans are clearly capable of performing at both ends of the spectrum and, thus, both facts must be accounted for in any workable theory. Although I dont have the answer, based on a range of empirical facts, I will try to spell out some of the properties I believe will be true of any theory of generic object recognition and categorization in humans.

References

- Kersten, D., Mamassian, P., and Yuille, A. (2004). Object Perception as Bayesian Inference. Annual Review of Psychology, 55, 271-304. http://gandalf.psych.umn.edu/~kersten/kersten-lab/papers/Kerstenannurevpsych.pdf

- Tarr, M. J. (2003). Visual Object Recognition: Can a Single Mechanism Suffice? In M. A. Peterson and G. Rhodes (eds.), Perception of Faces, Objects, and Scenes: Analytic and Holistic Processes (pp. 177-211). Oxford, UK: Oxford University Press. http://www.cog.brown.edu/~tarr/pdf/Tarr03.pdf

3.4. Dimensions of Neural Shape Space

C.E. Connor
John Hopkins University
USA

Abstract The visual system must somehow transform the extremely complex and variable retinal input image into a tractable, stable representation where object shape information is represented explicitly. Our studies of ventral pathway visual cortex suggest that the critical dimensions in the transformed representation relate to local contour properties: position (relative to other contour regions or to the object as a whole), orientation (1st derivative), and curvature (2nd derivative). Derivatives are useful for describing larger contour regions with fewer signals, and curvature can summarize contour regions large enough to be behaviorally and perceptually significant. Neurons in higher-level visual cortex span the position orientation curvature space with Gaussian-like tuning functions. A given contour region is represented by a population activity peak in this space. Whole objects are represented by multiple peaks corresponding to their constituent contour regions.

Refernces

- http://www.mb.jhu.edu/connor/media/pasupathy.pdf
- http://www.mb.jhu.edu/connor/media/pasupathy2001.pdf

3.5. Computational Mechanisms of Generic Object Recognition and Categorization in Cortex

Maximilian Riesenhuber
Department of Neuroscience
Georgetown University Medical Center
Washington, DC
USA

Abstract Object recognition is a difficult computational problem. Nevertheless, the human visual system can rapidly and effortlessly recognize objects in cluttered scenes under widely varying viewing conditions, at a level of performance far beyond that of current machine vision systems.

I will present a simple model of object recognition in cortex. The model, which is currently being used by a number of experimental and theoretical groups, accounts well for the complex visual task of object recognition in clutter, is biologically plausible, and makes nontrivial testable predictions. It consists of a hierarchy of processing stages based on just two different operations that serve to gradually increase shape specificity and invariance to stimulus transformations, producing a robust stimulus representation that permits the use of simple classifiers for recognition tasks.

I will then talk about experimental collaborations designed to test model predictions regarding (i) the neural mechanism underlying scale- and translation invariance and (ii) the neural bases of recognition tasks.

Finally, I will demonstrate the performance of the biological model using a benchmark face detection task on natural images. We find that the biological model performs as well or better than the comparison machine vision systems, and offers distinct computational advantages with respect to the complexity of the learning problem, transfer across different tasks, and invariance to scaling and translation.

References

- Riesenhuber, M., and Poggio, T. Neural mechanisms of object recognition. Current Opinion in Neurobiology 12, 162168 (2002).
- Serre, T., Riesenhuber, M., Louie, J., and Poggio, T. On the Role of Object-Specific Features for Real World Object Recognition in Biological Vision. In Biologically Motivated Computer Vision, Proceedings of Second International Workshop, BMCV 2002, Tubingen, Germany, November 22-24, 2002, Vol. 2525 of Lecture Notes in Computer Science, Springer, New York, 2002.
- http://riesenhuberlab.neuro.georgetown.edu

3.6. Representation of object images in the monkey inferotemporal cortex

Manabu Tanifuji
RIKEN Brain Science Institute 2-1 Hirosawa, Wako, Saitama 351-0198,
Japan

Abstract The monkey inferior temporal cortex (IT) is the association cortex implicated in object perception and recognition. Early studies showed that there are neurons responding to complex object images, such as faces, in this area. More recently, it has been also shown that many IT neurons respond to geometrically less complex features than to the more complex real objects.

Neurons responding to complex object images are not specific enough to a particular object image. For example, face neurons are not very selective to faces of different individuals. Similarly, the visual features represented by IT neurons are not complex enough to specify particular object images. Thus, in general, object images are represented by combined activation of these neurons. A question is how activities of these neurons are related to representation of object images.

To answer to the question, combination studies of functional imaging and single cellular recordings are useful. Functional imaging technique is advantageous to find multiple sites activated by an object image, and single cellular recordings enable us to characterize these sites in detail. These experiments showed that (1) local features of object images corresponds to some of the visual features represented by IT neurons, and (2) that some other visual features are related to global structures of object images, such as spatial relationship of parts. Face neurons responding arbitrary faces may be also related to global feature that is the configuration specific to faces.

References

- http://www.brain.riken.go.jp/english/b_rear/b1_lob/b1_7.html

3.7. End-to-End Learning of Object Categorization with Invariance Pose, Illumination, and Clutter

Yann LeCun, Fu Jie Huang
The Courant Institute, New York University
USA

Abstract We describe an end-to-end learning approaches to recognizing generic object categories with full invariance to pose, illumination, and clutter. The End-to-end learning approach consists in training the entire recognition system, from raw pixels to object categories, so as to minimize an overall discriminative performance measure.

A large dataset comprising stereo image pairs of 50 uniform-colored toys under 36 azimuths, 9 elevations, and 6 lighting conditions was collected (for a total of 194,400 individual images). The objects were 10 instances of 5 generic categories: four-legged animals, human figures, airplanes, trucks, and cars. Five instances of each category were used for training, and the other 5 for testing.

Low-resolution grayscale images of the objects with various amounts of variability and surrounding clutter were used to train and test Nearest Neighbor methods, and Support Vector Machines, operating on raw pixels or on PCA-derived features, and Convolutional Networks operating on raw pixels.

Experiments show that methods based on matching global templates (nearest neighbor and SVM) fare poorly with such a high intra-class variability. Convolutional nets, which are designed to learn a hierarchy of discriminative local features, yield error rates around 6.6% for classifying objects on a uniform background, 10.6% for detecting and classifying objects on textured background, and 16.7% on highly cluttered backgrounds. Experiments in monocular mode yielded considerably larger error rates, which suggests that convolutional nets can learn to take advantage of binocular inputs.

References

- http://www.cs.nyu.edu/~yann/

3.8. Toward True 3D Object Recognition

Jean Ponce
Beckman Institute and Department of Computer Science, University of Illinois at Urbana-Champaign
USA

Abstract This talk discusses two fundamental instances of the three-dimensional (3D) object recognition problem: (1) modeling rigid 3D objects from a small set of unregistered pictures and recognizing them in cluttered photographs taken from unconstrained viewpoints; and (2) representing object classes, learning the corresponding models from (possibly cluttered) sample pictures, and recognizing them in novel images despite clutter, occlusion, viewpoint and illumination changes, and individual variations within each class. I will review some of the main ideas underlying nearly 40 years of efforts at solving these two problems, identify some of the lessons learned along the way, and conclude with an overview of our current work, that revisits viewpoint invariants as a -local- representation of shape and appearance, and models 3D objects as collection of small (planar) patches, their invariants, and a description of their 3D spatial relationship.

Joint work with Svetlana Lazebnik, Frederick Rothganger, and Cordelia Schmid.

References

- see http://www-cvr.ai.uiuc.edu/ponce_grp/
- Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Local Affine Parts for Object Recognition. The Learning Workshop, Snowbird, Utah, 2004.
- Frederick Rothganger, Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Segmenting, Modeling, and Matching Video Clips Containing Multiple Moving Objects. Proc. CVPR'04.
- Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. A Sparse Texture Representation Using Local Affine Regions. Submitted to the IEEE Transactions on Pattern Recognition and Machine Intelligence, March 2004. (A preliminary version appeated in Proc. CVPR'03, Vol. II, pp. 319-324.)
- Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Affine-Invariant Local Descriptors and Neighborhood Statistics for Texture Recognition. Proc. ICCV'03, pp. 649-655.
- Fred Rothganger, Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. 3D Object Modeling and Recognition Using Affine-Invariant Patches and Multi-View Spatial Constraints. Proc. CVPR'03, Vol. II, pp. 272-277.

3.9. Recognizing Objects Using Shape

Jitendra Malik
U. C. Berkeley, CA
USA

Abstract TBD

References

3.10. A Shock-graph Dis-similarity Metric for Object Recognition

Benjamin B. Kimia
Division of Engineering
Brown University, USA
USA

Abstract The use of a suitable shape Representation is critical for a number of visual tasks. We describe how the shock graph, a dynamic hierarchical representation of the medial axis, is used for object recognition from silhouettes. Our approach is based on capturing the topology of the shape space via a dis-similarity metric that is the cost of the optimal deformation path between two shapes. Since the space of deformation paths is infinite-dimensional we discretize it by defining equivalence classes based on the shock graph topology and its transitions, which are related to the classical instabilities of the medical axis. A formal analysis of the local form of the shock graph and its transitions under a one-parameter family of deformations is described. The transition-based description of deformation paths is then searched under an edit-distance paradigm to find the optimal path. We describe recognition results which are stable under a range of visual transformations and which for several databases of up to 1032 shapes recover the correct category.

References

- http://www.lems.brown.edu/kimia.html

3.11. High-level Vision and Links to Language

David Forsyth
U. C. Berkeley, CA
USA

Abstract It is easy to forget that high-level vision is more than just template matching (or, for that matter, reasoning about geometric correspondence). Visual tasks humans can perform that are beyond the reach of current programs include: object recognition, where we identify instances of known objects despite vagaries of texture, geometry and view; object localization, where we determine where objects are with respect to one another, without necessarily knowing what the objects are; counting, which can again be done without knowing what objects are; and segmentation, where we identify where an object is in an image and in space without necessarily knowing what it is.

Many of these subtle and important tasks involve unknown or poorly understood objects. For example, we can determine how to grasp an object without identifying it. We can guess at a good path, and whether it will be dry or soggy underfoot. We can guess whether something provides a good handhold. We can guess whether to eat, ignore or flee from something without knowing precisely what it is. We can guess whether objects are heavy or light, wet or dry, rough or slippery, without knowing what they are. We can tell whether a predator is coiled to spring or snoozing without knowing much about its species or its behaviour. As a final example, we are capable of making a bewildering variety of deductions about other, unknown, individuals from relatively brief sightings of them moving around. Such activities involve a great deal more than efficient template matching.

I will discuss a variety of ways in which these problems have been reduced to fit the techniques of the day, covering my own work in geometric and statistical reasoning about matching. I will suggest that some hope of improved technique is to be obtained by considering both geometric and statistical together. Finally, I will point to language as a potential cue to interpreting some aspects of the visual world.

References

- http://www.cs.berkeley.edu/~daf/

4. Panel Discussion

The CVPR Workshop on Generic Recognition and Categorisation concluded with a panel session. It gave all the workshop participants, both speakers and the audience (more than 120 participants) a chance to engage in an open discussion. The panelists were Michael Tarr, Jitendra Malik, Gerard Medioni, and Max Riesenhuber, which nicely represented the overall interdisciplinary nature of the workshop (two computer vision panelists and two human vision panelists). The panel was chaired by Sven Dickenson.

A lively discussion developed around several issues:

- *Definition of a category*: There was a general agreement that there are different levels and types of categories among which many are not based on vision (or vision alone) but also on other criteria and modalities, e.g., affordances and function.

- *Categories based on local features*: Local features (in various variants capturing various invariances) are extensively used for categorizing objects in current systems. While such approaches are reasonably successful in cases when the set of categories is rather limited, it is clear that in more complex situations some structuring of the features is unavoidable. Perceptual grouping, segmentation, and categorization/recognition will have to be intertwined to arrive at some more complex features—parts. The issue of features/parts is also under extensive research in neurophysiology.

- *Categorization and learning*: The importance of learning was stressed since it is clear that a wide variety of possible features, parts, categories has to be obtained through a learning process. There was an interesting discussion on the number of training samples that are needed for learning. Since humans learn often from a rather small number of samples, this implies that there is much more (innate) structure than is usually assumed in computer vision.

- *Hierarchical representations*: This issue was raised by several participants since it is clear that features have to be organized over different levels of complexity. However, there has been no definite answer with respect to the number and granularity of these levels.

- *View-based versus object-centered representations*: There have been questions from the audience on a long-standing debate on view-based versus object-centered representations. Tarr claimed that it is a fruitless debate to engage in, namely, that there is evidence for both. Viewpoint dependence modulates with the particular recognition conditions. This is true regardless of the conditions under which the objects were

learned. Despite initial viewpoint invariance (consistent with some theories), object representations still include viewpoint dependent information.

- *Extensive testing*: To better understand the problems of categorization, there was a consensus that researchers should engage in extensive testing on large databases. These experiments should, among others, reveal intra-class variations of different categories and propose discriminant features, spatial structure of constellations of these features, etc. One such database is the one from Caltech, consisting of 101 categories.

Summary At the end of the panel, nobody claimed to have an answer to the categorization problem, which remains an open problem. Indeed, it is an incredibly hard problem, but we heard many interesting and thoughtful perspectives.