Learning objects and parts in images

Chris Williams, University of Edinburgh

This talk is in two parts. In the first part I will discuss our recent work on learning multiple objects/parts in images as an unsupervised learning (density estimation) problem. This builds on the framework of Jojic and Frey (CVPR, 2001) by using a robust statistical method that means that object models are extracted sequentially from the data. (Joint work with Michalis Titsias.) See http://www.dai.ed.ac.uk/homes/s0129556/lmo.html for further information.

In the second part of the talk I will describe work on hierarchical image modelling with dynamic trees (DTs). DTs are a tree-structured belief network framework where the configuration of the tree changes in response to the data, allowing the representation of different possible linkages of parts to wholes. (Joint work with Nick Adams, Steve Felderhof, Amos Storkey.)

Kernel methods and dimensionality reduction

Bernhard Schoelkopf

The talk will start with a short tutorial on kernel methods in machine learning. Following this, we will describe how some recent methods for nonlinear dimensionality reduction can be viewed as kernel methods.
http://www.kyb.tuebingen.mpg.de/publication.html?user=bs
http://www.kernel-machines.org

Object recognition as multimedia translation and data mining

Kobus Barnard, University of Arizona

I will present computer vision as a process that translates from visual representations (images) to semantic ones (words). This translation can be automatically learned from large unstructured data sets, suggesting that computer vision is a data mining activity focused on the relationships between data elements of different modes. More specifically, we link image regions to semantically appropriate words. Importantly, we do not require that the training data has the correspondence between the elements identified. For example, we may have keywords ”tiger” and ”grass” for an image, but we do not know whether the ”tiger” goes with a green region of the image, or the part with orange and black stripes. I will use an analogy with work in statistical machine translation for languages to explain how some of this ambiguity can be resolved with sufficient training data.

Replacing recognition with a similar but more easily characterized activity (word prediction) finesses many long standing problems. We do not prescribe in advance what kinds of things that are to be learned. This is an automatic function of the data, features, and segmentation. The system learns the relationships it can, and we do not have to construct by hand a new model in order to recognize a new kind of thing. Since we can measure system performance by looking at how well it predicts words for images held out from training, we use the system to evaluate image segmenters and choices of features in a principled manner. Finally, of great interest in our current work, the approach can be used to integrate high and low level vision processes. For example, we use word prediction to propose region merges. Using only low level features, it is not possible to merge the black and white halves of a penguin. However, if these regions have similar probability distributions over words, we can propose a region merge. If such a grouping leads to better overall word prediction, then it can be proposed as a (better) visual model for the word penguin.

Component-based object recognition
Bernd Heisele

I will present a component-based system for object detection. The system performs the detection by means of a two level hierarchy of classifiers. On the first level, the component classifiers independently detect parts of the object in the image. On the second level, a geometrical configuration classifier combines the results of the component classifiers and performs the final detection step. I will present techniques for determining a suitable set of components and discuss the importance of including position information about the detected components into the classification process. Finally, I will show applications of the component-based system to face detection, pedestrian detection, and face identification.

For publications see http://www.ai.mit.edu/projects/cbcl/publications/index-pubs.html

Using the forest to see the trees: a graphical model relating features, objects and scenes

Kevin Murphy, Antonio Torralba, Bill Freeman, MIT AI lab

Standard approaches to object detection try to classify local image regions, disregarding most of the rest of the image. Such local approaches may suffer from inherent ambiguities. In addition, it is computationally expensive to find the right local regions to classify, often requiring exhaustive search in position and scale. We show how using a low-dimensional summary of the whole image (the ”gist” of the scene) provides a cheap solution to both problems, since the gist can be easily computed, but provides a strong prior on what kinds of objects to expect, and where to expect them.

We develop a graphical model for combining scene priors with the output of various object detectors (based on boosting), and show improved results on a challenging data set of images collected from a wearable camera. We also briefly consider two extensions to the basic model. In the first extension, we perform joint recognition of many objects at once, using a conditional random field. In the second extension, we consider scene and object recognition over time, using a hybrid HMM model.

Invariant recognition of generic objects from shape

Yann LeCun

Yann LeCun, NYU, http://yann.lecun.com
Fu Jie Huang, NYU
Leon Bottou, NEC Labs America, http://leon.bottou.org

Detecting and recognizing generic objects independently of their location, scale, viewing angle, lighting condition, and surrounding clutter is a problem of considerable interest and considerable difficulty. Learning-based approaches to this problem have so far been rare in large part because of the the non-availability of a dataset with sufficient size and diversity to carry out meaningful experiments. We first introduce the NORB dataset which comprises stereo image pairs of 50 uniform-colored toys under 18 angles (every 10 degrees), 9 azimuths, and 6 lighting conditions, for a total of 97,200 individual images (images were collected every 10 degres, but every other angle was left out in the present experiments). The objects belong to 5 generic categories (with 10 instances for each): four-legged animals, human figures, airplanes, trucks, and cars. The raw images were used to generate very large datasets of 96x96 pixel greyscale stereo pairs with random variations of the position, scale, image-plane angles, luminosity, contrast, and background texture. Generic shape recognition experiments were run by training various models on 5 instances of each category, and testing on the remaining 5. Each category exhibits a very large variation of appearance, but contains essentially no usable feature except the shape (no color, no texture, and no distinctive local features).

Two training and testing sets were used for experiments. In the first set, called the *normalized-uniform dataset*, the objects were size-normalized, centered, and placed on uniform backgrounds. No random perturbation was performed. There was a total of 24,300 training samples and 24,300 test samples. In the second set, called the *jittered-textured dataset*, the objects were randomly shifted, scaled, rotated in-plane, modified for brightness and illumination, and placed on randomly chosen natural textures. A sixth class, containing only background textures (with no object) was added. There was a total of 291,600 training samples and 29,160 test samples.

Our aim is to assess how “standard” learning architectures can handle a task with such a high dimension and such large and complex intra-class variabilities. More specifically, we want to assess the extent to which “shallow”, template-based methods such as K-Nearest Neighbors and SVM can handle such problems, and to compare them with “deep” architectures such as convolutional networks. By shallow architecture, we mean one layer of template matchers followed by one layer of linear combinations. By deep architecture, we mean multiple trainable layers of non-linear transformations.

The test error rates on the *normalized-uniform dataset* were as follows: linear classifier: 30.2%; K-Nearest Neighbor (Euclidean distance): 18.4%; K-NN on 95 PCA features: 16.6%; SVM with Gaussian kernel on 32x32 monocular images 12.9%; SVM-Gaussian on 95 PCA features: 13.3%; convolutional net: 6.6%. First of all, it proved impractical to train an SVM on the full 96x96 pixel binocular images. This dataset is at the upper limit of practicality for K-NN and SVM because a template must be stored for each training sample. Second, even with no variation in the position, scale, and backgrounds, template-based methods are significantly worse than convolutional nets. Third, further experiments showed that the performance of convolutional nets is considerably better with binocular inputs than with monocular inputs.

Experiments were performed with linear classifiers and convolutional nets on the more realistic *jittered-textured dataset* (which includes 6 classes (5 object categories and background). Running SVM and K-NN on a dataset this large is beyond the reasonnable. The test error rate was 30.6% for linear classifiers and 7.1% for convolutional nets (much of the errors were between the truck and car categories). These numbers provide a baseline for future work on learning-based methods for object recognition from shape with invariance to pose, lighting, and background textures. Improved performance can be expected from methods that incorporate a bit more prior knowledge about image formation. A live demo of real-time object recognition with convolutional nets will be shown.

Learning a rare event detection cascade by direct feature selection

Jim Rehg, Georgia Tech

Face detection is a canonical example of a rare event detection problem, in which target patterns occur with much lower frequency than non-targets. Out of millions of face-sized windows in an input image, for example, only a few will typically contain a face. Viola and Jones recently proposed a cascade architecture for face detection which successfully addresses the rare event nature of the task. A central part of their method is a feature selection algorithm based on AdaBoost. We present a novel cascade learning algorithm based on forward feature selection which is two orders of magnitude faster than the Viola-Jones approach and yields classifiers of similar quality. This faster method could be used for more demanding classification tasks, such as on-line learning or searching the space of classifier structures. Our experimental results highlight the dominant role of the feature set in the success of the cascade approach.

This is joint work with Jianxin Wu and Matthew D. Mullin.

Approximate nearest neighbors: beyond the L_p norms

Jeff Erickson, UIUC (posing as Piotr Indyk, MIT)

This talk will survey both classical and recent results on finding approximate nearest neighbors in high-dimensional Euclidean and other metric spaces. Specific topics will include kd-trees and their recent variants, locality-sensitive hashing, and quasi-isometric embeddings of Hausdorff distance and earth-mover's distance.

Geometric algorithms for biomedical applications

Carola Wenk, University of Arizona

There is a broad range of geometrical problems which naturally arise in life science areas. This talk will survey a few of them which deal with comparing geometric shapes ('shape matching'). I will give a short introduction into this area, and focus on presenting two biomedical projects that I am involved in:

"Protein Gel Matching": The growing field of proteomics deals with identifying and analyzing the functions of proteins. This is a very hot topic in molecular biology. In order to get hold of certain proteins a protein probe is usually separated by 2D gel electrophoresis yielding a 2D image of axis-parallel ellipses each representing a single protein. The first task ("spot detection") is to identify the ellipses, which turns into a geometric covering task when ellipses overlap. The comparison of two 2D electrophoresis gels ("gel matching") is an essential tool to identify variations of the amount of certain proteins present in a probe. Our matching algorithm detects the spots in two gel images in a preprocessing step, and then operates on point patterns. For the gel matching we use tools from computational geometry, and obtained the first "automatic" matching algorithm that is currently used in a commercial software product.

"Brain Matching": In neurosurgery the surgeon has to be careful not to destroy important parts of the brain. There are modern tools which allow the surgeon, during the surgery, to point at a position in the brain which is then displayed in a precomputed 3D image of the patient's brain. This requires a robust registration between the 3D image and the brain during the surgery. I will present the algorithm we developed for this task, which matches a set of markers that are attached to the patient's head.

See http://www.cs.arizona.edu/people/carolaw/cvCW.html for online available publications, and http://gelmatching.inf.fu-berlin.de for a demo and more information on the gel matching program.

Object recognition in the geometric era : a retrospective

J. L. Mundy, Brown University

We are embarked on a new era in object recognition where intensity and reflectance play a more central role in the formulation of features. This talk will summarize the several decades of recognition research that focused on purely geometric representations. The goal of the presentation is to revisit the basic assumptions of geometry-based recognition, and the outcome of a mature line of research, to see what can be learned that might impact the way forward.

Object class recognition by unsupervised scale-invariant learning

Rob Fergus

We present a method to learn and recognize object class models from unlabeled and unsegmented cluttered scenes in a scale invariant manner. Objects are modeled as flexible constellations of parts. A probabilistic representation is used for all aspects of the object: shape, appearance, occlusion and relative scale. An entropy-based feature detector is used to select regions and their scale within the image. In learning the parameters of the scale-invariant object model are estimated. This is done using expectation-maximization in a maximum-likelihood setting. In recognition, this model is used in a Bayesian manner to classify images. The flexible nature of the model is demonstrated by excellent results over a range of datasets including geometrically constrained classes (e.g. faces, cars) and flexible objects (such as animals).

Web page: http://www.robots.ox.ac.uk/ vgg/publications/html/index.html

Fragment-based recognition and segmentation

Shimon Ullman

I will describe an approach to visual classification and segmentation, which is based on representing shapes within a class by a combination of shared sub-structures called fragments. The fragments are sub-images selected from a training set of images, based on a criterion of maximizing the mutual information of the fragments and the class they represent. They are then used to represent a variety of different object views within a given class. Segmentation and recognition are performed in this scheme together, in contrast to the more traditional view in which image segmentation is performed first, in a bottom-up manner, followed by object recognition.

Borenstein, E. and Ullman, S. (2001) Class specific top down-segmentation. Proceedings of the European Conference on Computer Vision, 110-122.

Ullman, S., M. Vidal-Naquet, and S. Sali, (2002) Visual features of intermediate complexity and their use in classification. Nature Neuroscience, 5(7)), 1-6.

Qualitative shape matching for object and action recognition

Stefan Carlsson

Recognition of object classes poses problems of shape and appearance variation that are much harder than those encountered in the recognition od specific objects. We will present an algorithm for matching perceptually corresponding shapes based on representing order properties of the image geometry. Using ordinal as opposed to metric structure captures qualitative shape properties that typically define object and action classes and it allows for a wider range of variation in the data to be matched. In the first part we will present results in recognition of object and action classes based on an algorithm for computing point to point correspondence between shape prototypes and the image to be recognized. In the second part we will present preliminary results on direct shape matching of images by computing order properties from the image gradient field.

Object recognition at GE

Anthony Hoggs

The Computer Vision Group at GE Research is currently engaged in the development of technology in medical imaging, industrial inspection, broadcast media and video surveillance for various GE businesses and the US government. In all of these applications, object recognition is an important technology. For GE Medical Systems, we are developing algorithms to automatically detect early-stage cancer in volumetric images. For GE Aircraft Engines, we are developing a system to detect minute manufacturing errors in jet turbine blades by comparing range imagery to detailed 3D models. For NBC Television, we are performing research for automatically annotating broadcast news with semantic video content such as objects, people and events. For GE Industrial Systems, we are conducting research in recognizing arbitrary objects in retail video surveillance data. For Lockheed Martin, we are investigating the recognition of vehicles in aerial video.

These areas and many other potential GE applications would benefit from fundamental advances in object recognition. In particular, the research topics of most interest are: using multiple views and/or multiple video frames; exploiting prior, approximate geometry; exploiting prior semantic knowledge or context; hierarchical, categorical recognition; and handling low resolution, noisy images.

http://www.research.ge.com/vision/

France Telecom's expectations and research in object recognition
Henri Sanson, Christophe Laurent, Olivier Bernier

In this presentation, we first recall the main markets of France Telecom and the increasing importance of audio-visual contents for all these activities. We then explain why we believe that object recognition will be inescapable in two main applications, namely audio-visual content indexing and retrieval, and Human Service Interaction, for providing powerful and attractive telecommunication-related services. We finally present a brief overview of our current research relating to still and moving image analysis, indexing and object recognition [FBVC00], [FBVC01], [VB02], [LLB03],[LLV03],[DLL03].

References

[DLL03] Dorval T., Laurent C., and Laurent N., *A Friendly Object Selection Tool Combining Salient Features Detection and Potential-Based Active Contour*, Proceedings of the Third International Workshop on Content-Based Multimedia Indexing (CBMI'03) (Rennes (France)), September 2003.

[FBVC00] R. Feraud, O. Bernier, J.E. Viallet, and M. Collobert, *A fast and accurate face detector for indexation of faces images*, FG'2000 Conference on Automatic Face and Gesture Recognition (Grenoble, France), March 2000.

[FBVC01] ______, *A fast and accurate face detector based on neural networks*, IEEE Transactions on pattern analysis and machine intelligence **23** (2001), no. 1.

[LLB03] Laurent C., Laurent N., and Bodo Y., *A Human Skin Detector Combining Mean Shift Analysis and Watershed Algorithm*, Proceedings of IEEE International Conference on Image Processing (ICIP'03) (Barcelona), September (Spain) 2003.

[LLV03] Laurent C., Laurent N., and Visani M., *Color Image Retrieval Based on Wavelet Salient Features Detection*, Proceedings of the Third International Workshop on Content-Based Multimedia Indexing (CBMI'03) (Rennes (France)), September 2003.

[VB02] J.E. Viallet and O. Bernier, *Face detection for video summaries*, CIVR'2002 the Challenge of Image and Video Retrieval (London), July 2002.

3D object recognition at Toyota

Yutaka Hirano, Toyota Motor Corporation

Toyota is now developing 3D object recognition technologies for dexterous hand robot. Our main focuses are on 1. detection of object from cluttered background, 2. separating objects which occludes each other. 3.Pose estimation of objects and 4. 3D reconstruction of environment and making map for locomotion. I will introduce Toyota's needs and approach for 3D object recognition for the robot.

Invariant local features for recognition

David Lowe, Cordelia Schmid, Jiri Matas, Tinne Tuytelaars

Local photometric features have become popular as a practical and effective approach to image matching and recognition. They are distinctive as well as robust to occlusion and clutter. Recently these features have been extended to be scale and affine invariant which allows matching and recognition in the presence of large scale and viewpoint changes. We present these features and describe their application to image matching, 3D recognition, robot navigation and construction of image panoramas from unordered sets of images as well as texture classification and detection of object categories.

Topics:

Interest point detectors and descriptors
- Scale invariance
- Distinctive local descriptors
- Affine invariance
- Experimental comparison of detectors and descriptors

Applications of local descriptors
- Wide baseline image matching
- Object recognition and robot localization
- Construction of image panoramas from unordered sets of images
- Texture classification and detection of object categories

Links to recent papers on this topic:
http://www.cs.ubc.ca/ lowe/papers/ijcv03-abs.html
http://www.cs.ubc.ca/spider/lowe/papers/brown03-abs.html
http://www.inrialpes.fr/lear/people/schmid/cvpr-tutorial03
http://cmp.felk.cvut.cz/ matas/papers/matas-bmvc02.pdf
ftp.esat.kuleuven.ac.be/pub/psi/visics/tuytelaa/papers/Tuytelaars_VanGool-wbs_air-subm_IJCV.pdf

Texture recognition using affine-invariant regions

Svetlana Lazebnik, University of Illinois

This talk will discuss texture representations using affine-invariant interest points. A model of a texture is constructed from a sparse set of image locations characterized by local appearance and affine shape. For more descriptive power, it is possible to incorporate neighborhood constraints based on co-occurrence statistics. Applications include retrieval, classification, and segmentation of images of textured surfaces under a wide range of transformations, including viewpoint changes and non-rigid deformations.

References:

Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce, “Affine-Invariant Local Descriptors and Neighborhood Statistics for Texture Recognition,” to appear in ICCV 2003.

Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce, “A Sparse Texture Representation Using Affine-Invariant Regions,” CVPR 2003, vol. II, pp. 319-324.

Papers are available at http://www-cvr.ai.uiuc.edu/ponce_grp/

Exploring images for object recognition

Luc Van Gool

Object recognition systems based on local invariant features have recently become increasingly popular. Even if successful for a range of problems, the robustness and applicability of these approaches are limited by the repeatability of the feature extraction and difficulties when matching due to clutter, substantial scale changes, very oblique viewing, serious occlusions, etc. We try to remedy these problems by abandoning strategies of relying solely on the matching of individual invariant regions. Instead, we use an initial set of seed matches to then gradually explore the surrounding area, attempting to construct new matches. The augmented match set is then pruned by a topological filter. The system goes through several such cycles of expansion and contraction, From cycle to cycle, more matches are found and the percentage of correct matches in the set increases. Apart from the increased robustness of the process, it offers other advantages, like the ability to deal with deforming objects, and the ability to come up with the approximate outline of the visible parts of the object as a side-product.

Recognition by parts

Martial Hebert

We present an overview of recognition systems based on parts indexing for the recognition of objects in images. The parts are image patches represented by local histograms of filters, contours, and intensity distribution. The key aspect of the work is the development of efficient techniques for indexing. In particular, a distance measure that is efficient in discriminating between classes of objects based on parts content is defined. Efficient approaches to indexing based on this distance measure are also defined. Examples are given on different recognition tasks, such as the identification of desktop objects and face identification. The approach is compared with other common indexing techniques used in recognition. Comparisons using standard face datasets and other datasets are discussed.

Application of Fisher information to line detection

Steve Maybank, University of Reading

The framework for many detection problems includes a parameterised family of probability density functions $\theta \rightarrow p(x|\theta)$ where θ is a parameter and x is a measurement. The task is to find a value of θ compatible with a given set S of measurements. Examples of such detection problems include the detection of lines, circles and ellipses in images, the detection of the epipolar transform between pencils of epipolar lines and the detection of the collineation between two images of a plane.

An asymptotic approximation is obtained for the Fisher information of the family of densities $\theta| \rightarrow p(x|\theta)$. Under the Fisher information the parameter space becomes a Riemannian manifold. These results are the basis of a new algorithm for detecting lines in images. The parameter space for line detection embeds isometrically as a surface of revolution in the Euclidean space $R^{(}3)$. Lines are detected by sampling the parameter space at a finite number of points and checking each sample point for compatibility with the measurements.

http://www.cvg.cs.reading.ac.uk/ sjm/

Illumination and reflectance modelling, and its application to face recognition
David Kriegman

Over the past decade there have been significant advances in understanding the relation between lighting, reflectance, and the set of images of an object under different lighting conditions. While this understanding has lead to new reconstruction algorithms and rendering techniques, it also has implications for object recognition. While it has been shown that the set of images of an object in fixed pose under all lighting conditions is a convex cone in the space of all images, earlier empirical evidence showed that this set can in fact be approximated by a low-dimensional linear subspace. In the last few years, these empirical results have been explained using spherical harmonics to represent both lighting and reflectance. These theoretical developments have been applied particularly in the domains of face tracking, face recognition, and image clustering. This talk surveys both these results on modeling the effects of lighting and the application area of face recognition.

Image parsing by data-driven markov chain monte carlo

Song-Chun Zhu, UCLA

Image parsing refers to the task of decomposing an image into its constituent patterns: texture/color/shading regions, curves and curve groups, and objects etc. This task subsumes traditional vision problems such as segmentation, grouping, and recognition. It entails searching a very complex state space for Bayesian optimization.

In this talk I shall review some recent results in our group for designing rapid mixing Markov chains to explore complex solution spaces, such as DDMCMC for segmentation, Swendsen-Wang Cut for graph partition, these designs are found to be effective in various subspaces. I will also show various analyses on the MCMC speed. These analyses explain why data-driven (heuristics) can expedite the Markov chain search.

I shall also address the relationship between generative models and discriminative models, and their roles in visual knowledge representation and computing.

http://civs.stat.ucla.edu/Segmentation/Segment.htm

Learning from observation : object and task recognition for programming by demonstration

Katsushi Ikeuchi

We have been developing a programming-by-demonstration system to generate robot programs through observing human actions. The main idea is not simply to mimic human hand trajectories such as tele-operations or motion-based systems, but to understand each human action through generating symbolic state-representations and describing such action by their transitions, and then to map such state transitions to robot motions. We consider that this state-based system is more robust and flexible than traditional motion-based systems. In our state-based system, the first step of analysis is to recognize objects and to extract ”relations” among objects; the second step is to obtain relation transitions; the third step is to match such relation transitions with abstract task models, pre-compiled and stored in a computer, to infer actions to generate such transitions. We refer this framework as task recognition as a parallel to object recognition. In this talk, we will explain three task recognition systems; the first system recognizes operations of polyhedral objects using face contact transitions. The second system recognizes operations of mechanical parts using mechanical links among parts. Third system recognizes operations of strings using the knot theory and the Reidemister moves.

Posters

3D object modeling and recognition using affine-invariant patches and multi-view spatial constraints

Fred Rothganger, Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce

This work presents a novel representation for three-dimensional objects in terms of affine-invariant image patches and their spatial relationships. Multi-view constraints associated with groups of patches are combined with a normalized representation of their appearance to guide matching and reconstruction, allowing the acquisition of true three-dimensional affine and Euclidean models from multiple images and their recognition in a single photograph taken from an arbitrary viewpoint. The proposed approach does not require a separate segmentation stage and is applicable to cluttered scenes. Preliminary modeling and recognition results are presented.

Discriminative random fields for modeling spatial dependencies in images

Sanjiv Kumar and Martial Hebert, CMU

In this work we present Discriminative Random Fields (DRF), a discriminative framework for the classification of natural image regions by incorporating neighborhood spatial dependencies in the labels as well as the observed data. The proposed model exploits local discriminative models and allows to relax the assumption of conditional independence of the observed data given the labels, commonly used in the Markov Random Field (MRF) framework. The parameters of the DRF model are learned using penalized maximum pseudo-likelihood. Furthermore, the form of the DRF model allows the MAP inference for binary classification problems using the graph min-cut algorithms. The performance of the model was verified on the synthetic as well as the real-world images. The DRF model outperforms the MRF model in the experiments.

SVM-based nonparametric discriminant analysis, an application to face detection

Rik Fransens, Tinne Tuytelaars, Luc Van Gool

Detecting the dominant normal directions to the decision surface is an established technique for feature selection in high dimensional classification problems. Several approaches have been proposed to render this strategy more amenable to practice, but they still show a number of important shortcomings from a pragmatic point of view. This presentation introduces a novel such approach, which combines the normal directions idea with Support Vector Machine classifiers. The two make a natural and powerful match, as SVs are located nearby and fully describe the decision surfaces. The approach can be included elegantly into the training of performant classifiers from extensive datasets. The potential is corroborated by experiments, both on synthetic and real data, the latter on a face detection experiment. In this experiment we demonstrate how our approach can lead to a significant reduction of CPU-time, with neglectable loss of classification performance.

Scene categorization by learning repeated elements

Mark Everingham, University of Oxford

Images of both natural and man-made scenes contain repeated elements at a number of scales. We explore whether groupings of these repeated elements can be used as a basis to classify images and label image regions.

We develop a statistical model based on grouped repeated elements. The parameters of the model are learnt from training images in order to classify images of various types of scenes. Grouping is involved at two stages: first, within an image, where repeated elements are grouped despite changes in appearance and shape; and second, between images, where the grouping must be unaffected by intra-class variation. An unseen image can then be classified by the class-conditional probability of the groups discovered in the image.

Shape warping and statistics

Guillaume Charpiat, ENS

Differentiable approximations of shape metrics (such as the Hausdorff distance) provide a simple way to warp a shape onto another by solving a Partial Differential Equation (PDE). Their first order variation defines a normal deformation field for a given curve, from which we can define the mean of several shapes, their covariance ”operator”, and the principal modes of variation.

Demos

Wide-baseline matching and object recognition using extremal regions
Jiri Matas

A wide baseline matching and recognition systems exploiting the Maximally Stable Extremal Regions will be demonstrated. The systems implements strategies for establishing correspondences based both on rotational invariants and on matching of local affine frames.

References:

Jiri Matas, Ondrej Chum, Martin Urban, and Tomas Pajdla. Robust wide baseline stereo from maximally stable extremal regions. In Paul L. Rosin and David Marshall, editors, Proceedings of the British Machine Vision Conference, volume 1, pages 384-393, London, UK, September 2002. BMVA.
http://cmp.felk.cvut.cz/ matas/papers/matas-bmvc02.pdf.

Stepan Obdrzalek and Jiri Matas. Local affine frames for image retrieval. In Michael S. Lew, Nicu Sebe, and John P. Eakins, editors, CIVR'02: Proceedings of International Conference The Challenge of Image and Video Retrieval, volume 1, pages 318-327, Berlin, Germany, July 2002. Springer-Verlag.
http://cmp.felk.cvut.cz/ matas/papers/obdrzalek-civr02.pdf.

Video google: a text retrieval approach to object matching in videos

Josef Sivic and Andrew Zisserman, University of Oxford

We will demonstrate an approach to object and scene retrieval which searches for and localizes all the occurrences of a user outlined object in a video. The object is represented by a set of viewpoint invariant region descriptors so that recognition can proceed successfully despite changes in viewpoint, illumination and partial occlusion. The temporal continuity of the video within a shot is used to track the regions in order to reject unstable regions and reduce the effects of noise in the descriptors.

The analogy with text retrieval is in the implementation where matches on descriptors are pre-computed (using vector quantization), and inverted file systems and document rankings are used. The result is that retrieval is immediate, returning a ranked list of key frames/shots in the manner of Google.

The demo will include matching on two full length feature films ('Run Lola Run' and 'Groundhog Day').