**Figure 2:**
**Focal attention at regions where spatial statistics break with the generally observed statistics in natural scenes, that is, at regular patterns.**

visual input into a plurality of invariant representations is a necessary information-reduction stage in any cognitive vision system.

To limit the enormous computational burden arising from the complex task of interpretation and learning, any efficient general vision system will ignore the common statistics in its input signals. Hence, the apparent occurrence of invariant representations decides what is salient and therefore requires attention. Such focal attention is a necessary selection mechanism in any cognitive vision system, critically reducing both the processing requirements and the complexity of the visual learning space, and effectively limiting the interpretation task. Expectation about the scene is then inevitably used to steer attention selection. Hence, focal attention is not only triggered by visual stimuli, but is affected by knowledge about the scene, initiating conscious behaviour. In this principled way, knowledge and expectation may be included at an early stage in cognitive vision. In the near future, we intend to study the detailed mechanisms behind such focal attention mechanisms.

**Links:**
http://www.ecvision.info/home/Home.htm,
http://www.science.uva.nl/~mark/

**Please contact:**
Jan-Mark Geusebroek and Arnold W.M. Smeulders, University of Amsterdam
Tel: +31 20 525 7552
E-mail: {mark, smeulders}@science.uva.nl

# Coarse-to-Fine Object Detection

**by François Fleuret and Hichem Sahbi**

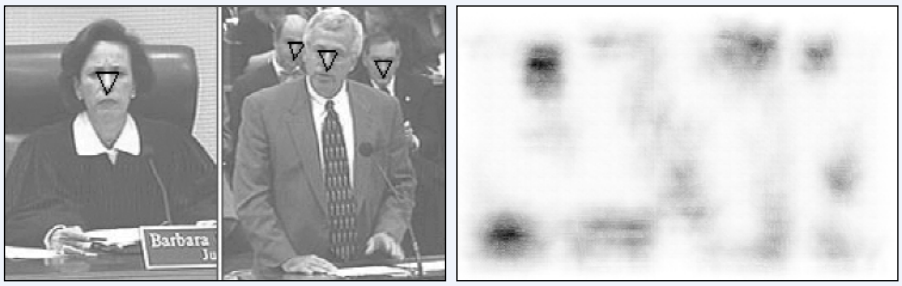
**Of all the techniques currently available, object recognition remains one of the most promising approaches to content-based image retrieval. The advent of software able to detect and automatically recognise a wide range of objects in pictures would lead to a new generation of technologies based on high-level semantics. Customers could, for instance, be provided with interactive on-line catalogues, improving the search capabilities in TV network archives, which contain usually hundreds of thousands of hours of video.**

In the classical approach, object detection is related to the broader field of pattern classification with statistical methods. Given a large number of examples, those techniques build classifiers that are able to label small patches of scenes as object or non-object. Such a classifier is used on an entire scene, and at all possible scales, to achieve object detection. As it can be expected, putting aside issues related to learning, such a brute-force approach necessarily has a very high computational cost. This is an important drawback, since the major areas using detection, such as real-time detection in video or indexing of large databases of pictures, require very low computation times in order to process several scenes per second.

INRIA's IMEDIA research team is studying a family of algorithms which explicitly address the trade-off between error rate and computational cost. We have developed several detectors based on the same idea of a hierarchical composition of classifiers, which reflects a hierarchical decomposition of the views of the object to be detected. Each one of these complex detectors is an adaptive sequence of tests: at each step, a classifier is chosen according to the responses received so far, and its own response is then computed.

This global approach has several important advantages. Firstly, it concentrates the computation on ambiguous areas; trivial areas are rejected early in the process, after a very small computation investment. The second advantage is that it dispatches the representation of the object being detected to a large number of learning machines, which can be very simple. For instance, this approach does not expect the classifiers to learn invariance to geometrical transformations of pictures, as such transformations are taken into account at the global level of the detector. The approach is also generic in respect of the type of classifiers. We have used either very simple edge-counting schemes or more sophisticated wavelet-based support vector machines.

Given a large set of classifiers of varying cost and statistical power, we have



**The figure on the left shows the result of face-detection. Each triangle indicates the location of a detected face, with an accurate pose estimation. The figure on the right gives a graphical representation of the computation intensity on the various parts of the pictures. Intensity is estimated by counting the number of times the algorithm takes into account the value of each pixel during the processing of the scene. White stands for a couple of such access, while the dark areas correspond to several hundreds of them.**

studied how they can be combined in order to obtain the optimal average cost when processing a scene. This optimisation is based on a statistical model of the relation between the cost and the error rate of the individual classifiers. This leads to a constrained optimisation problem that can then be solved. As might be intuitively expected, both power and cost must grow, so the process begins with simple, almost trivial tests, and goes into details later. More precisely, the growth in complexity is exponential. The resulting strategy is able to reject trivial parts of the picture - like an empty sky or a wall - with very little computational cost (see Figure 1).

Future work will address the handling of large families of objects. Because it dispatches representations to a large number of classifiers, we expect the coarse-to-fine approach to enable control of the growth of both representation and computation when dealing with larger families of objects.

**Links:**
http://www-rocq.inria.fr/~sahbi/Web/these.html
http://www-rocq.inria.fr/~fleuret/ctf.html

**Please contact:**
Hichem Sahbi, INRIA
Tel: +33 1 3963 5870
E-mail: Hichem.Sahbi@inria.fr

François Fleuret, INRIA
Tel: +33 1 3963 5583
E-mail: Francois.Fleuret@inria.fr

# Video Understanding and Indexing for Surveillance: Image Perception, Quality and Understanding

**by Tamás Szirányi**

**Motion tracking and scene analysis, especially in surveillance systems, require a high-level interpretation of possible shapes and their events, even in the case of incomplete vision conditions and transient motion. To this end, a multi-camera surveillance system has been developed and new efficient algorithms constructed at the Analogical and Neural Computing Systems Laboratory of SZTAKI, with fragments of motion and bodies being grouped through methods of statistical inference.**

What is the common thread in the indexing of archive films and the registration of objects in surveillance tasks? In both cases, some understanding of the scene is required, and in the case of films we can be sure that the director's intention lies behind the movement of the camera. In surveillance, however, there is no director, area of focus or even structured scene. Statistical inference is needed to extract meaningful information when it is required. Clever effects such as illumination or associations can help to elucidate incomplete visual data to support human visual understanding. These effects are exploited in composed films, and could be also helpful in interpreting poor surveillance data.

If surveillance problems are related back to better-defined scenes, then Bayesian approaches can lead to an understanding of unanticipated events. The main problem involves the duration of events. While film sequences have a beginning and an end, surveillance events are usually transient dissolving scenes. The arbitrary motions of low-resolution objects (think of a police camera surveying a whole street) are not easily interpreted. However, if we compose beforehand a set of possible events with motion samples and statistical inductions, then real-life surveillance events can be more easily analysed. While it is an achievement to successfully connect the two distinct areas of visual analysis, problems can occur on both sides: definitions of objects and motion in transient events, indexing of sequences and interpretation of scenes.

Motion and shape can hardly be described in real-life applications of noisy video surveillance. Using a greater number of