# Cascaded attention and grouping for object recognition from video*

**C. Greindl, A. Goyal, G. Ogris, and L. Paletta†**

Computational Perception Group
Institute of Digital Image Processing, JOANNEUM RESEARCH
Wastiangasse 6, A-8010 Graz, Austria

## Abstract

*Object detection is an enabling technology that plays a key role in many application areas, such as content based media retrieval. Cognitive vision systems are here proposed where the focus of attention is directed towards most informative processing. The attentive detection system uses a cascade of increasingly complex classifiers of radial basis functions (RBF) networks for the stepwise identification of regions of interest (ROIs) and refined object hypotheses. While the coarse classifier is used to determine first approximations on the ROI, more complex classifiers are used to give sufficiently accurate and consistent pose estimates. Objects are modelled by local appearances and in terms of posterior distributions in eigenspace. The experimental results were led for the automatic detection of brand objects in Formula One broadcasts and clearly illustrate the benefit in applying decision making on attention and probabilistic grouping.*

## 1 Introduction

The upcoming of content based media retrieval, ambient intelligence systems, miniaturised mobile sensors and requirements on visual security systems imposes additional challenges on the robustness and accuracy of *object detection* systems. Despite the fact that there exist impressive sample systems for face and people detection (e.g., [5]), more complex issues such as 3-D object detection, partial occlusion, and maintenance of quality of service remain to be thoroughly treated.

Attentive cognitive vision systems are here proposed where the focus of attention is directed towards the most relevant information or the most promising operation for the task at hand. Information is integrated in a sequential process that dynamically makes use of knowledge and that enables spatial grouping on the local appearances. Cascaded classification [16] is in this context understood as multi-step information extraction, focusing attention within each classification step on recursively refined features of a given problem. Attentive information selection, grouping, and detection of spatial context is defined by constraints on the geometry of spatial relations between local object appearances which are powerful indicators for discrimination. Attention mechanisms have been recently reported being essential for object recognition (e.g., [3]) and in processing cascades [16]. Grouping processes are known to occur across the field of view [3, 10], and interrelate to a mid level abstraction of object representations [10]. Grouping is a way of binding local appearances to form a structure of spatial context, as a basis unit of attention [3].

In computer vision, attentive processing was initially based on the saliency of responses of feature detectors which contribute to the locating of matching candidates [13]. More elaborated models of visual search were proposed [4] which first combine multiscale image features into a single topographical saliency map. Competition among neurons in this map gives rise to a single winning location that corresponds to the next attended target in a purely stimulus-driven manner. A model visual attention based on the concept of selective tuning is outlined by Tsotsos et al. [15] The central thesis is that attention acts to optimise the search procedure inherent in a solution to vision. The system aspect of behavior based information selection has been investigated in recent models that focus on decision making aspects in vision tasks [2].

The original contribution of the presented work is to clearly demonstrate superior performance of a general attention framework that integrates hierarchical classifiers with spatial information fusion, producing evi-

† *lucas.paletta@joanneum.at*

Proceedings of the 12th International Conference on Image Analysis and Processing (ICIAP'03)
0-7695-1948-2/03 $17.00 © 2003 **IEEE**

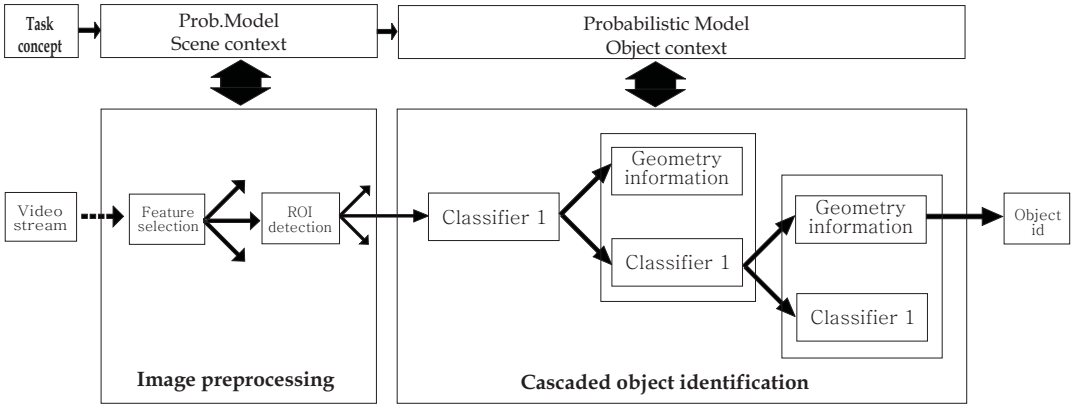



**Figure 1. Concept of attentive object detection. The cascaded object detection system applies decision making to focus attention on information sources and actions. From a coarse classification the system either attends to a more refined classification stage or to a structural pose evaluation of local information chunks.**

dence that multiple information sources together with cognitive vision systems are mandatory for robust and accurate image processing.

The proposed attentive object detection system is outlined using a cascade of increasingly complex classifiers for the stepwise rejection of object hypotheses and pose estimates (Fig. 1). While the most simple classifiers are used to determine first approximations on a region of interest (ROI) in the input image [16], more complex classifiers are used for more refined ROIs to give more confident estimates. Thereby, this method presents an element of choice as a component of a global decision making control (e.g., [8]). Finally, we apply a probabilistic framework on all observable variables, and, in particular, to model the local spatial context. Objects are modelled by posterior distributions from local appearances [11] modelled by probabilistic radial basis function (RBF) networks.

For given results of a ROI, the decision is on either to select a more complex classifier in order to get more confident object hypotheses or to perform a grouping to extract pose structure from topological constraints. This attentive decision making focuses computing resources on different actions and information sources.

The experiments were led for the automatic detection of brand objects within 'Formula One' videos.

## 2 Object based attentive grouping and classification

In this section, we describe individual stages of the object identification module illustrated in Fig. 1 (cascaded object identification). A region of interest is first classified using a rapid but coarse posterior neural classifier (Section 2.1). Based on the consistency of a region's classifications, the detection system then either applies a full pose estimation from the topological relations within the ROI (Section 2.2), or uses a more refined but complex neural classifier (Section 2.1) in order to receive more accurate predictions, and to group consistently interpreted regions.

### 2.1 Object priming using RBF cascaded neural classifiers

Probabilistic radial basis functions (RBF) networks [9] are here used to represent stages of a classification cascade that recursively refines its object posterior estimates on a given ROI. The initial classification stage minimizes network complexity under maximization of classification accuracy. Subsequent stages aim at increasing precision by allowing more complex neural architectures.

In this work we assume a local receptive field approach that represents an object part, i.e., the imagette [11], by a vector in eigenspace derived from principal component analysis (PCA). Variation of a visual parameter $\varphi_j$ (e.g., imagette pose within object $o_i$) determines the object's *trajectory model* in eigenspace. Recently, the requirement to formulate object representations on the basis of local information has been broadly recognized [11, 6], e.g., for increased tolerance to partial occlusion, improved accuracy of recognition (since only relevant - i.e., most discriminative - infor-

Proceedings of the 12th International Conference on Image Analysis and Processing (ICIAP'03)
0-7695-1948-2/03 $17.00 © 2003 **IEEE**

mation is queried for classification), etc.

Probabilistic RBF networks (see [8]) apply a Bayesian framework with density estimations provided by unsupervised clustering, where the confidence estimates are refined by supervised learning. The eigenspace feature vector $\mathbf{y}$ is fed to the RBF network and mapped to the output values $z_k$ for a posterior estimate, $\hat{P}(O_k|\mathbf{y}) = \alpha z_k(\mathbf{y})$, $\alpha$ is a normalizing constant. A decision on object recognition is accordingly applied using a Maximum A Posteriori (MAP) decision.

### 2.2 Grouping from topological information

The relevance of structural dependencies in object representation [11, 12] has been stressed before, though the existing methodologies merely reflect co-location in the existence of local features. The presented work outlines full integration of geometrical relations between local features within a framework on Bayesian conditional analysis of *perception-action* sequences, that perform a grouping like behavior, as follows.

A Bayesian analysis of context is here defined on the basis of the local statistics about the visual information. Given the measurement about object $o_i$ under visual parameter $\varphi_j$, the likelihood of obtaining feature vector $\mathbf{y}$ is denoted by $p(\mathbf{y}|o_i, \varphi_j)$. The likelihood might be estimated from a set of sample images with fixed $o_i$ and $\varphi_j$, capturing the inaccuracies in the parameter $\varphi_j$ such as moderate light variations or segmentation errors [8]. Via Bayesian inversion one obtains then $P(o_i, \varphi_j|\mathbf{y}) = p(\mathbf{y}|o_i, \varphi_j)P(\varphi_j|o_i)P(o_i)/p(\mathbf{y})$. A posterior estimate with respect to the object hypotheses $o_i$ is given by $P(o_i|\mathbf{y}) = \sum_j P(o_i, \varphi_j|\mathbf{y})$.

Geometrical information is derived from the relation between the stored object model - the set of trajectories in feature space - and the actions (shift between receptive fields on object image) that are mapped to changes in the pose parameter $\varphi_j$. Introducing the representation of actions $a_i$ into Bayesian fusion leads to

$$P(o_i, \varphi_j|\mathbf{y}_1, a_1, \mathbf{y}_2) = \alpha P(o_i, \varphi_j|\mathbf{y}_1, a_1)p(\mathbf{y}_2|o_i, \varphi_j, \mathbf{y}_1, a_1). \quad (1)$$

Spatial context is now exploited using the conditional term $P(o_i, \varphi_j|\mathbf{y}_1, a_1)$: The probability for observing view $(o_i, \varphi_j)$ as a consequence of deterministic action $a_1 = \Delta\varphi_1$ must be identical to the probability of having measured at the action's starting point before, i.e. at view $(o_i, \varphi_j - \Delta\varphi_1)$, thus $P(o_i, \varphi_j|\mathbf{y}_1, a_1) \equiv P(o_i, \varphi_j - \Delta\varphi_1|\mathbf{y}_1)$.

Furthermore, the probability density of $\mathbf{y}_2$, given the knowledge of view $(o_i, \varphi_j)$, is conditionally independent on previous observations and actions, and therefore $p(\mathbf{y}_2|o_i, \varphi_j, \mathbf{y}_1, a_1) = p(\mathbf{y}_2|o_i, \varphi_j)$. The recursive update rule for *conditionally dependent* observations accordingly becomes,

$$P(o_i, \varphi_j|\mathbf{y}_1, a_1, \ldots, a_{N-1}, \mathbf{y}_N) = \alpha p(\mathbf{y}_N|o_i, \varphi_j)P(o_i, \varphi_j - \Delta\varphi_{N-1}|\mathbf{y}_1, a_1, \ldots, \mathbf{y}_{N-1}) \quad (2)$$

and the posterior, using $\mathbf{Y}_N^a \equiv \{\mathbf{y}_1, a_1, \ldots, a_{N-1}, \mathbf{y}_N\}$, is then given by $P(o_i|\mathbf{Y}_N^a) = \sum_j P(o_i, \varphi_j|\mathbf{Y}_N^a)$.

The experimental results in Figures 2 and 3 demonstrate that context is crucial for rapid discrimination from local object information. The grouping of conditionally observable variables to an entity of semantic content, i.e., a visual object, is an essential perceptual process [10, 3]. This section evaluates geometrical information from configurations of local responses.

## 3 Experimental Results

The experiments were performed on object detection in 'Formula One' broadcast videos. Outdoor events provide challenging data due to illumination changes, changes in scale and pose, and partial occlusions. The industry is interested in application such as annotation of sport events [1] or automatic detection of company brands (logos) for publicity evaluation.

The proposed object detection system is schematically performing according to Fig. 1,

1. A coarse RFBN classifier is applied within the initial ROI information.

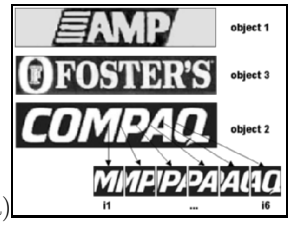
2. From regions of consistent object hypothesis attribution, one extracts the geometry from local information, i.e., the spatial context from a set of individual imagettes (Section 2.2), and determines object poses thereby.

3. For inconsistent regions, a more complex RBF classifier is applied so that the pose evaluation of this remaining part becomes sufficiently reliable (should only contain information from one object).

4. From updated consistent regions, one extracts the geometry from local information for pose estimation.

5. Remaining small gaps of inconsistent object attribution are either operated by even more complex classificators, or denoted non-object regions, or are interpolated to fit to a globally consistent object pose attribution.

(a)




(b)

| param. | | 1-step recognition [%] | | | | N-step recognition [%] | | | |
|---|---|---|---|---|---|---|---|---|---|
| Δf | Δs | AMP | COM | FOS | avg. | AMP | COM | FOS | avg. |
| 2 | 2 | 91,9 | 71,0 | 41,1 | 68,0 | 96,8 | 83,9 | 47,6 | 76,1 |
| 5 | 5 | 92,2 | 75,7 | 45,6 | 71,2 | 99,0 | 98,1 | 99,0 | 98,7 |
| 10 | 20 | 100,0 | 82,1 | 14,3 | 65,5 | 96,4 | 92,9 | 92,9 | 94,0 |

**Table 1. (a) Logos used in the experiments, with imagettes expanded for context fusion. (b) Evaluation of the accuracy in object recognition for different numbers of fusions and shift steps (actions), and 3 objects (AMP, COM, FOS - see (a).**

| | Deviation from true pose | | |
|---|---|---|---|
| Method | ±0 [%] | ±2 [%] | ±5 [%] |
| 1-step (2,2) | 0,5 | 1,1 | 11,9 |
| N-step (5,5) | 83,0 | 98,7 | 99,0 |

**Table 2. Evaluation of the accuracy in pose recognition for $\Delta f = 2/\Delta s = 2$ and $\Delta f = 5/\Delta s = 5$ respectively (see Table 1), and different tolerance to pose variations.**

### 3.1 Experiments on object priming from RBF networks

Object identification from local context is applied on an initially extracted ROI, assuming that the video frames have been segmented from learned color filters [7]. Each object imagette sample projects into eigenspace of a particular dimension. This vector is then input to a probabilistic RBF neural classifier of a certain computational complexity, majorly determined by the size of the hidden layer. Depending on the size of the hidden layer and the number of eigendimenions representing the input, the recognition accuracy can vary distinctively. In the experiments, we used for the coarse pre-classification an RBF network with 100 hidden units and 3 eigendimensions, and for the refined object classification 400 hidden units and 18 eigendimensions. Note that these classifications, where the MAP confidence was beyond a threshold confidence value (e.g., 0.9), naturally exhibit a high level of accuracy on the test data.

Spatial context from geometry can be easily extracted based on a predetermined estimate on scale and orientation of the object of interest (Tables 1 , 2, Fig, 2). This is computed (i) from the geometry of the ROI, and from (ii) estimates on pose and scale from global image features [14, 7].

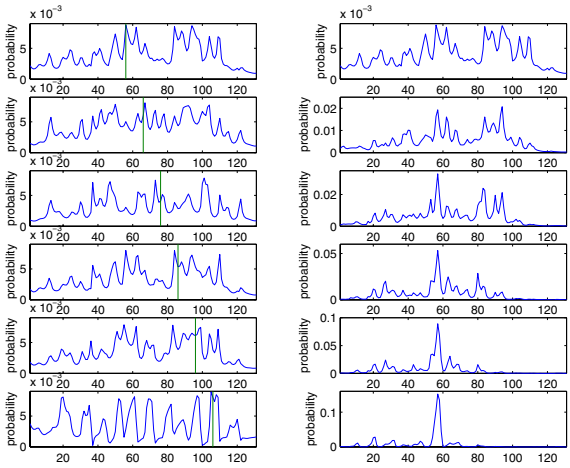



**Figure 2. Probability distributions over pose hypotheses (imagette pose no. 1-127 within logo 'COM') from individual test imagettes no. 1-6, from top to bottom.** *Right:* **Corresponding fusion results using spatial context from geometry illustrating fusion steps no. 1-5.**

### 3.2 Experiments on attentive object and pose estimation

Fig. 3b-e illustrates the experiments on attentive mechanisms for the object and pose recognition of the "AMP" logo. Fig. 3a shows a frame with a sample object, and corresponding evaluations of pose (see below on figure: "nf"=n-tuple fusion, "1f"=no fusion, "Obj-id nf"=object identitty from nf, "detmap"=detection map; green=correct pose/object, pink=wrong). A first scan of the coarse RBF classifier (see first paragraph above) results in a region that provides consistent object hypotheses with high confidences (Fig. 3b,c - see

Proceedings of the 12th International Conference on Image Analysis and Processing (ICIAP'03)
0-7695-1948-2/03 $17.00 © 2003 **IEEE**

poses no. 1-40). However, a majority of the poses (no. 41-127) required an updated object classification so that consistent regions could be identified therein (Fig. 3e,f) as well. Note that the pose estimation demonstrated highly accurate results across the entire logo course (Fig. 3g), due to the attentive treatment of ROI information.

## 4 Conclusions

Attention mechanisms are an essential part of video interpretation systems that require robust object detection. This paper describes the concept and the early stages of an object detection system with a methodology that understands attention in terms of goal driven decision making and concerted behavior based recognition. A first stage of information selection is applied from extracting the object specific context of simple features to provide an indexing into regions of interest. Decision making and information selection is then operated on the ROI dependent on the consistency of coarsely classified parts of the region. Only reliable segmentations are fed into a pose estimator based on local spatial context. Regions that are less reliably interpreted require more refined recognition stages as induced by more complex RBF neural classifiers. The preliminary experiments demonstrate that it obviously makes sense to apply the attentive scheme of cascaded object classifiers in order to find out consistent regions and enable therefore reliable information fusion patterns.

Future work will first evaluate extensive experiments on real world data and aims at achieving a video rate implementation of this method. Secondly, a probabilistic modelling of the complete processing chain needs to be outlined, from the early stages of ROI detection to grouping process and, finally, Bayesian object identification.

## References


[1] J. Assfalg, M. Bertini, C. Colombo, and A. Del Bimbo. Semantic annotation of sports videos. *IEEE Multimedia*, 9(2):52–60, 2002.

[2] C. Bandera, F. J. Vico, J. M. Bravo, M. E. Harmon, and L. C. Baird. Residual Q-learning applied to visual attention. In *Proc. International Conference on Machine Learning*, pages 20–27, 1996.

[3] G.W. Humphreys. A multi-stage account of binding in vision: Neuropsychological evidence. *Visual Cognition*, 8:381–410, 2001.

[4] L. Itti and C. Koch. Computational modelling of visual attention. *Nature Reviews Neuroscience*, 2:1–10, 2001.

[5] A. Mohan, C. Papageorgiou, and T. Poggio. Example-based object detection in images by components. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(4):349–361, 2001.

[6] S. Obdrzalek and J. Matas. Object recognition using local affine frames on distinguished regions. In *Proc. British Machine Vision Conference*, 2002.

[7] L. Paletta and Christian Greindl. Context based object detection from video. In *Proc. International Conference on Computer Vision Systems*. Graz, Austria, in print, 2003.

[8] L. Paletta and A. Pinz. Active object recognition by view integration and reinforcement learning. *Robotics and Autonomous Systems*, 31(1-2):71–86, 2000.

[9] T. Poggio and S. Edelman. A network that learns to recognize three-dimensional objects. *Nature*, 317:314–319, 1990.

[10] R. A. Rensink. The dynamic representation of scenes. *Visual Cognition*, 7:17–42, 2000.

[11] B. Schiele and J.L. Crowley. Recognition without correspondence using multidimensional receptive field histograms. *International Journal of Computer Vision*, pages 31–50, 2000.

[12] C. Schmid. A structured probabilistic model for recognition. In *Proc. IEEE International Conference on Computer Vision*, 1999.

[13] M. Swain and D. Ballard. Color indexing. *International Journal of Computer Vision*, 7(1):11–32, 1991.

[14] A. Torralba and P. Sinha. Statistical context priming for object detection. In *Proc. IEEE International Confernce on Computer Vision*, 2001.

[15] J. K. Tsotsos, S. M. Culhane, W. Y. K. Wai, Y. Lai, N. Davis, and F. Nufl. Modeling visual attention via selective tuning. *Artifical Intelligence*, 78:507–545, 1995.

[16] P. Viola and M. Jones. Rapid object detection using a bossted cascade of simple features. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2001.



Proceedings of the 12th International Conference on Image Analysis and Processing (ICIAP'03)
0-7695-1948-2/03 $17.00 © 2003 **IEEE**

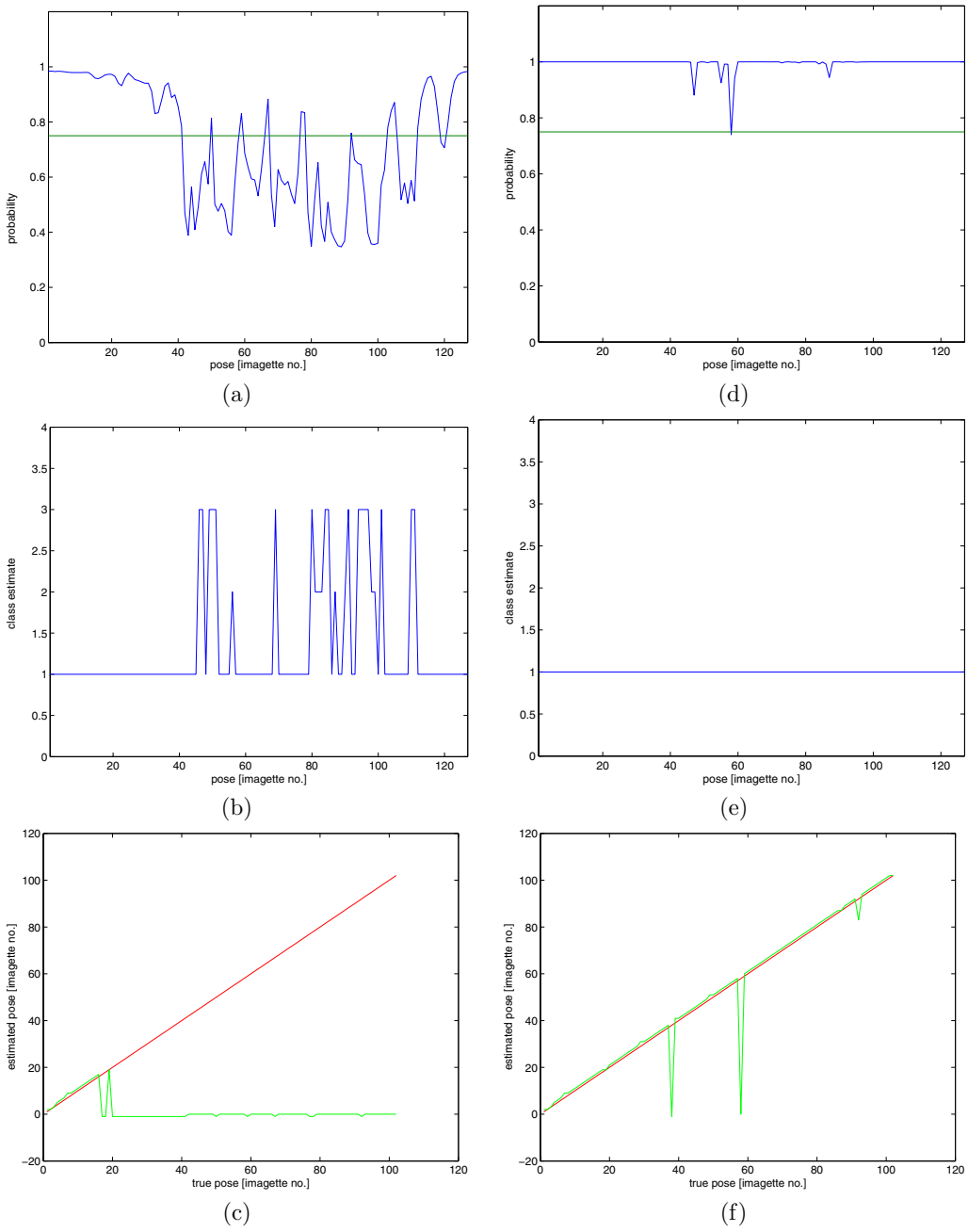



**Figure 3. Object identification and pose estimation with 3-dimensional (a-c) and 18-dimensional (d-f) eigenspace input. (a,d) illustrates the confidence value over imagette poses of the classified object hypothesis based on a posterior MAP decision. (b,e) denotes the classification results (classes 1-3), (c,f) plots the expected vs. the estimated object pose.**

Proceedings of the 12th International Conference on Image Analysis and Processing (ICIAP'03)
0-7695-1948-2/03 $17.00 © 2003 **IEEE**