

# Learning to Focus Attention on Discriminative Regions for Object Detection

Gerald Fritz and Christin Seifert and Lucas Paletta<sup>1</sup> and Horst Bischof<sup>2</sup>

**Abstract.** A major task of visual attention is to focus processing on regions of interest to enable rapid and robust object search. Instead of integrating generic feature extraction into object specific interpretation we strictly pursue a top-down approach. Early features are tuned to selectively respond to task related visual features, i.e., locally discriminative information that is useful in object recognition. In this work we determine discriminative regions from the information content in the local appearance patterns. A rapid mapping from appearances to discriminative regions is estimated using decision trees. The focus of attention on discriminative patterns enables then efficient detection of a searched object, but also the definition of sparse object representations to respond only to task relevant information. In the experiments, the performance in object recognition from single images dramatically increased considering only discriminative patterns. Evaluation of complete image analysis under various degrees of partial occlusion and image noise resulted in highly robust recognition even in the presence of severe occlusion and noise effects. Finally, preliminary results on attentive object detection in cluttered environments demonstrated successful indexing to relevant locations.

## 1 INTRODUCTION

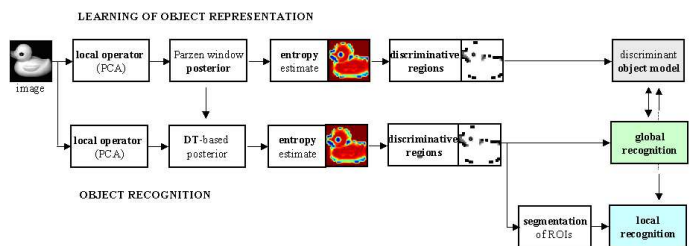
Cognitive computer vision systems must address detection and recognition of objects in the on-going stream of visual experience. A major issue is to decide about at which level of bottom-up signal interpretation should top-down information affect recognition processing. Recognition from early local information may serve several purposes, such as, improved tolerance to occlusion effects, or to provide initial evidence on object hypotheses in terms of providing starting points in cascaded object detection.

Previous work on the exploitation of local operator responses for attentive object search developed several viewpoints on how to advance from local to global information.

In object recognition from local information, [13] applied standard interest operators (Förstner, Harris) with the aim to determine localizable object parts for further analysis. In addition, [13] selected class related patterns from related fixed spatial configurations of recognizable operator responses. To avoid dependency on scale selection, [5, 6, 8] introduced interest point detectors that derive scale invariance from local scale saliency. These operators proved to further improve recognition from local photometric patterns, such as in [3]. While concern has been taken specifically with respect to issues

of scale invariance [5], wide baseline stereo matching performance [6, 8], or unsupervised learning of object categories [13, 1, 3], the application of interest point operators has not yet been investigated about the information content they provide with respect to object discrimination.

The key contribution to visual attention is to investigate information theoretic saliency measures with respect to object search and recognition. This work firstly provides an analysis of the discriminative power of local appearances, and secondly, exploits discriminative object regions to build up an efficient local appearance representation and recognition methodology (Fig. 1). In contrast to the use of



**Figure 1.** Concept of the entropy based attentive detection model. (Top) From a subspace feature selection (e.g., PCA) we compute the posteriors and the local information content. Responses of discriminative regions with entropy values below a predefined threshold are stored in the object model. (Bottom) For rapid object recognition, a decision tree outputs local entropy estimates. A global recognition method operates on the complete image. In contrast, individual discriminative regions can be segmented to provide a local recognition decision.

classifiers that determine discriminant features (e.g., [12]) for recognition, our approach intends to make the actual local information content explicit for further processing, such as, constructing the object model (Sec. 2) or determining discriminative regions for recognition and detection (Sec. 3).

We propose in a first stage to localize discriminative regions in the object views from the Shannon entropy of a locally estimated posterior distribution (Sec. 2.1). In a second stage, we consequently derive object models in feature subspace from discriminative local patterns (Sec. 2.2). Object recognition is then exclusively applied to test patterns with associated low entropy. Identification is achieved by majority voting on a histogram over local target attributions (Sec. 3). Rapid object recognition is supported by decision tree bases focus of attention on discriminative regions of interest (Sec. 4). The method is evaluated on images degraded with Gaussian noise and different degrees of partial occlusions using the COIL database (Sec. 5).

<sup>1</sup> Joanneum Research, Institute of Digital Image Processing, Wastiangasse 6, A-8010 Graz, Austria, email: lucas.paletta@joanneum.at

<sup>2</sup> Graz University of Technology, Institute for Computer Graphics and Vision, Inffeldgasse 6, A-8010 Graz, Austria, email: bischof@tugraz.ac.at

## 2 ENTROPY-BASED OBJECT MODELS

The proposed object model consists of projections of those local appearances that provide rich information about an object identity, i.e., *reference imagettes*<sup>3</sup> mapped into a subspace of the corresponding image matrix. Local regions in the object views that are both discriminative and robustly indicate the correct object label provide the reference imagettes for the object representation.

### 2.1 Local distributions in subspace

We use a principal component analysis (PCA, [7]) calculated on local image windows of size  $w \times w$  to form the basis for our local low dimensional representation. PCA maps imagettes  $\mathbf{x}_i$  to a low dimensional vector  $\mathbf{g}_i = \mathbf{E}\mathbf{x}_i$  by matrix  $\mathbf{E}$  consisting of few most relevant eigenvectors of the covariance matrix about the imagette sample distribution  $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_i, \dots, \mathbf{x}_N\}$ ,  $N$  is the total number of imagette samples.

In order to get the information content of a sample  $\mathbf{g}_i$  in eigenspace with respect to object identification, we need to estimate the entropy  $H(O|\mathbf{g}_i)$  of the posterior distribution  $P(o_k|\mathbf{g}_i)$ ,  $k = 1 \dots \Omega$ , where  $\Omega$  is the number of instantiations of the object class variable  $O$ . The Shannon entropy denotes

$$H(O|\mathbf{g}_i) \equiv - \sum_k P(o_k|\mathbf{g}_i) \log P(o_k|\mathbf{g}_i). \quad (1)$$

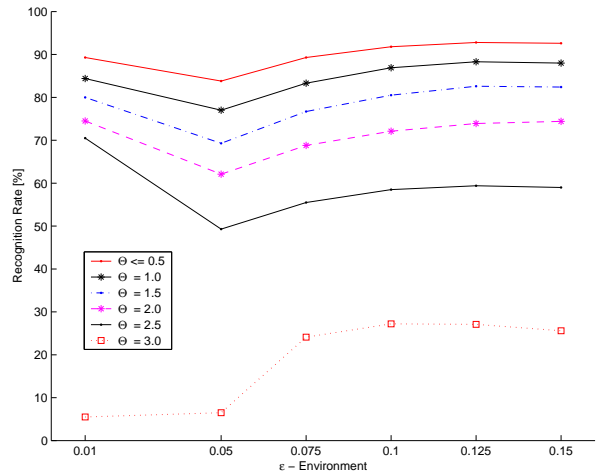
We approximate the posteriors at  $\mathbf{g}_i$  using only samples  $\mathbf{g}_j$  inside a Parzen window [10] of a local neighborhood  $\epsilon$ ,  $\|\mathbf{g}_i - \mathbf{g}_j\| \leq \epsilon$ ,  $j = 1 \dots J$ . We weight the contributions of specific samples  $\mathbf{g}_{j,k}$  - labelled by object  $o_k$  - that should increase the posterior estimate  $P(o_k|\mathbf{g}_i)$  by a Gaussian kernel function value  $\mathcal{N}(\mu, \sigma)$  in order to favour samples with smaller distance to observation  $\mathbf{g}_i$ , with  $\mu = \mathbf{g}_i$  and  $\sigma = \epsilon/2$ . The estimate about the Shannon entropy  $\hat{H}(O|\mathbf{g}_i)$  provides then a measure of ambiguity in terms of characterizing the information content with respect to object identification within a single local observation  $\mathbf{g}_i$ .

### 2.2 Discriminative object regions

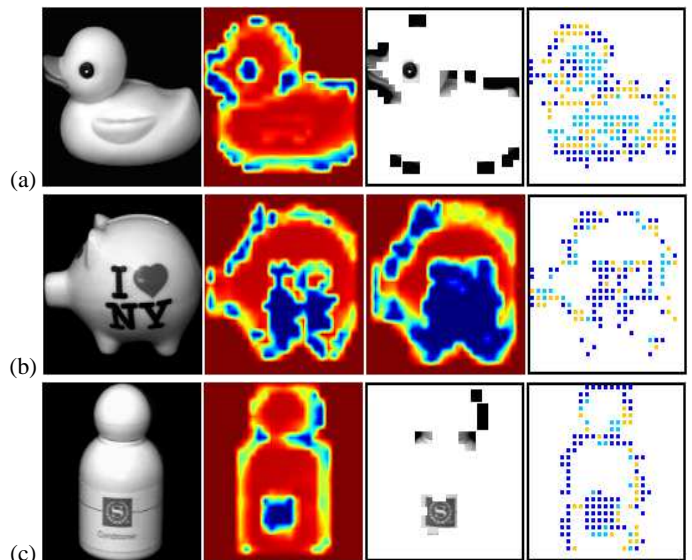
It is obvious that the size of the local  $\epsilon$ -neighborhood (Sec. 2.1) will impact the distribution and thereby the recognition accuracy (Fig. 2, Sec. 5), highly depending on the topology of object related manifolds in subspace. One can construct an entropy-coded image of an object view from a mapping of local appearances  $\mathbf{x}_i$  to corresponding entropy estimates (Fig. 3 for characteristic objects of the COIL-20 database, Sec. 5). The individual images (from left to right) depict the original image, the color coded entropy image (from  $9 \times 9$  pixel imagettes), corresponding imagettes with  $\hat{H}(O|\mathbf{g}_i(\mathbf{x}_i)) \leq 0.5$ , and associated images coding recognition accuracy (blue=correct, white=false). From these images it becomes obvious that regions containing specific texture and brightness contrasts provide highly discriminative information for recognition.

From discriminative regions we proceed to *entropy thresholded object representations*. The entropy coded images provide evidence that segmentation into discriminative regions and consequently exclusive usage of the associated reference points in eigenspace would provide sparse instead of extensive object representations [2], in

<sup>3</sup> imagettes denote subimages of an object view [2]



**Figure 2.** Recognition performance using MAP classification on samples of a neighborhood  $\epsilon$ . Rejecting imagettes with entropy  $\hat{H}(O|\mathbf{g}_i) > \Theta$ ,  $\Theta = 2.0$ , may dramatically increase accuracy of overall object recognition.



**Figure 3.** Sample COIL-20 objects (a)  $o_{11}$ , (b)  $o_{13}$ , (c)  $o_{16}$  at view 0; each with - left to right - (i) original frame, (ii) entropy image (from  $9 \times 9$  pixel imagettes; entropy coded color palette from low=blue to high=red), (iii) local appearances with  $\Theta \leq 0.5$  in a,c (entropy image from  $15 \times 15$  pixel imagettes in b), and (iv) accuracy-coded images (accuracy blue=true, white=false).

terms of storing only imagette information that is *relevant for classification* purposes. Object representations from local photometric patterns have been constructed either from extensive storage of all subspace (reference) points for k-nearest neighbor matching [2], or from selected, cluster specific prototype points [13, 3] that necessarily convey uncertainty. In contrast, the proposed object model includes only *selected* reference points for nearest neighbor classification, storing exclusively those  $\mathbf{g}_i$  with

$$\hat{H}(O|\mathbf{g}_i) \leq \Theta. \quad (2)$$

A specific choice on the threshold  $\Theta$  consequently determines both storage requirements and recognition accuracy (Sec. 5). To speed up the matching we use efficient memory indexing of nearest neighbor candidates described by the adaptive  $K$ - $d$  tree method [4].

### 3 OBJECT RECOGNITION FROM LOCAL INFORMATION

The proposed recognition process is characterized by an entropy driven selection of image regions for classification, and a voting operation, as follows,

1. **Mapping** of imagette patterns into subspace (Sec. 2.1).
2. **Probabilistic interpretation** to determine local information content (Eq. 1).
3. **Rejection** of imagettes contributing to ambiguous information (Sec. 2.2).
4. **Nearest neighbor analysis** of selected imagettes within  $\epsilon$ -environment.
5. **Majority voting** for object identifications over a full image nearest neighbor analysis.

Each imagette pattern from a test image that is mapped to an eigenspace feature point is analyzed for its entropy  $\hat{H}(O|\mathbf{g}_i)$  with respect to object identification. In case this imagette would convey ambiguous information, its contribution to a global recognition decision would become negligible, therefore it is removed from further consideration. Actually, practice confirms the assumption that it is difficult to achieve a globally accurate object identification when multiple ambiguous imagettes ‘wash out’ any useful evidence on a correct object hypothesis [9]. The entropy threshold  $\Theta$  for rejecting ambiguous test points in eigenspace is easily determined from the corresponding threshold applied to get the sparse model of reference points by Eq. 2. Selected points are then evaluated on whether they lie within the  $\epsilon$ -distance of any model reference point. In case several points are identified, the object class label of the nearest neighbor point is associated to the queried test point.

Object recognition on a set of imagettes is then performed on finding the object identity by majority voting on the complete set of class labels attained from individual imagette interpretations.

### 4 ATTENTIVE OBJECT DETECTION FROM LOCAL INFORMATION

For the purpose of *rapid* object recognition and detection, we need a mapping of low computational complexity to perform a *focus of attention* on regions of interest (ROIs). These ROIs would then be fed into the recognition module (Sec. 3) for detailed analysis. Actually, this kind of segmentation would work in terms of a point of interest operator (POI) and would be tuned by a discriminative objective function (entropy based). In order to keep complexity of the POI low, we do not use a neural network approach or a universal function estimator. In contrast, we provide rapid entropy estimates from a decision tree classifier, assuming appropriate quantization of the entropy values into class labels. One expects from this tree to provide this estimate from a few attribute queries which would fundamentally decrease computation time per image for ROI computation.

**Estimation of entropy values** For a rapid estimation of local entropy quantities, each imagette projection is fed into the decision tree which maps eigenfeatures  $\mathbf{g}_i$  into entropy estimates  $\hat{H}$ ,  $\mathbf{g}_i \mapsto \hat{H}(\Omega|\mathbf{g}_i)$ . The C4.5 algorithm [11] builds a decision tree using the standard top-down induction of decision trees approach, recursively partitioning the data into smaller subsets, based on the value of an attribute. At each step in the construction of the decision tree, C4.5 selects the attribute that maximizes the information gain ratio.

The induced decision tree is pruned using pessimistic error estimation [11].

**Rapid extraction of ROIs** The extraction of ROIs in the image is performed in 2 stages. First, the decision tree based entropy estimator provides a rapid estimate of local information content. Only eigenfeatures  $\mathbf{g}_i$  with an associated entropy below a predefined threshold  $\hat{H}(O|\mathbf{g}_i) < H_\Theta$  are considered for recognition (Sec. 3). These selected discriminative eigenfeatures are then processed by nearest neighbor analysis with respect to the object models and by majority voting according to the process described in Sec. 3.

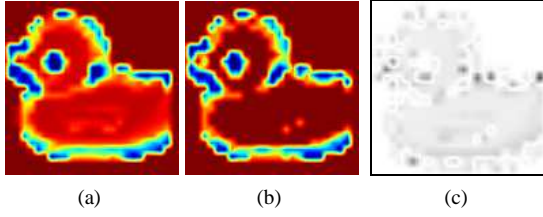
## 5 EXPERIMENTAL RESULTS

In order to perform a thorough analysis of the object recognition performance we applied the described methodology to images of the COIL-20 database [7].

**Eigenspace representation** Experiments were applied on 72 (test: 36) views of 20 objects of the COIL-20 database, for a total of 1440 (720) gray-scaled images with normalized size of  $128 \times 128$  pixels. Analysis was performed with various imagette sizes:  $9 \times 9$  (selected for further experiments),  $15 \times 15$ , and  $21 \times 21$  pixels, resulting in larger discriminative regions (normalized per entropy maximum; e.g., in Fig. 3b) for larger imagette sizes. However, recognition errors due to partial occlusion positively correlate with imagette size. Imagettes were sampled by a step size of 5 pixels, giving a total of 603950 (301920) imagettes for training (test), excluding black background imagettes. Imagettes were projected into a 20-dimensional eigenspace.

**Local information content** For a local probabilistic interpretation of test imagettes, we searched for an appropriate threshold  $\epsilon$  to determine the training samples in a local neighborhood (Sec. 2.1) that will be considered for the posterior estimation. Fig. 2 shows recognition rates (using a MAP classifier and images degraded with 10% Gaussian noise, operated on all *single* imagettes) with various values for neighborhood  $\epsilon$  and entropy threshold  $\Theta$ . This diagram shows that imagettes with high entropy  $\Theta > 2$  dramatically decrease the recognition rate.  $\epsilon = 0.1$  was selected for best performance while not taking too many imagettes into account.

**Discriminative regions from decision trees** The decision tree was trained using eigenfeatures  $\mathbf{g}_i$  of 50% of all extracted imagettes from the COIL-20 database and associated entropy estimates, determined from the Parzen window approach (Sec. 2.1). Entropy values were linearly mapped into equally spaced  $N_H$  intervals  $[(k-1)H_{max}/N_H, kH_{max}/N_H]$  for  $k = 1..N_H$ ,  $N_H = 5$  (Tab. 1). The error on the training set was determined 2.4%, the error on the test set 13.0%. This suffices to map eigenfeatures very efficiently to corresponding entropy intervals (classes). Fig. 4 illustrates a typical decision tree based estimation of local entropy values. It demonstrates that discriminative regions can be represented highly accurately (Fig. 4b) from sequences of only  $\approx 25$  attribute queries. A difference image (Fig. 4c) reflects the neglectable errors and confirms that discriminative regions can be both reliably and rapidly estimated from C4.5 decision trees.



**Figure 4.** Estimation of entropy from decision trees 4(a) Entropy estimation from Parzen windows  $I_P$  (Sec. 2.1, color code see Fig. 3), 4(b) entropy estimation from decision tree  $I_T$  (Sec. 4, 4(c) difference image  $I_D = I_P - I_T$  (grayvalue coded for  $[0, H_{max}]$  from white to black)

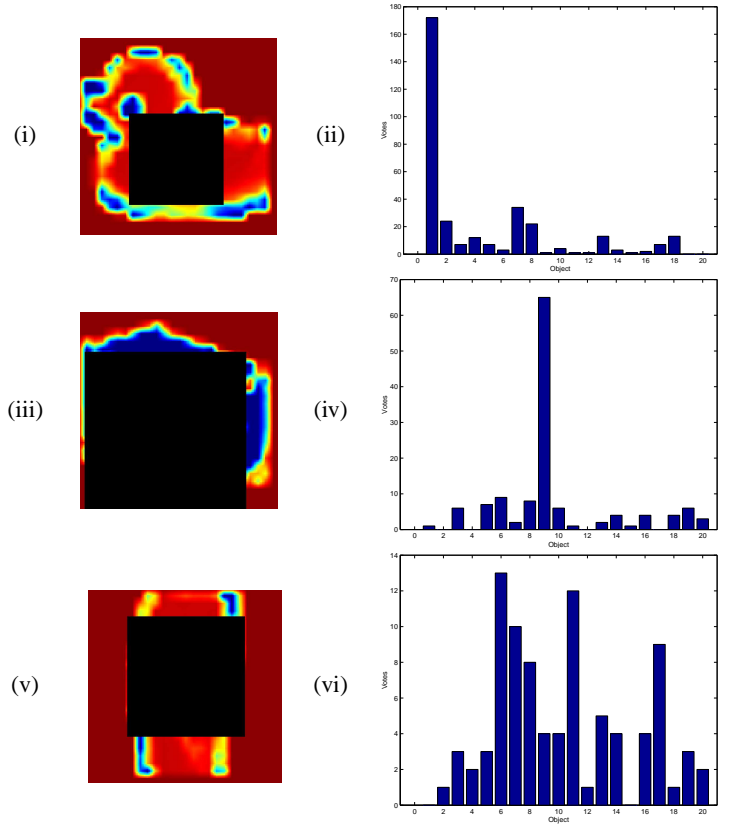
maps $\mapsto$	c1	c2	c3	c4	c5
c1	38146	5546	2981	1425	218
c2	1820	44418	1382	624	72
c3	1495	2094	41989	2548	190
c4	1036	1136	3725	40775	1644
c5	166	192	400	2709	44849

**Table 1.** Confusion map of the C4.5 decision tree based entropy estimation. The individual entropy intervals - denoted by classes  $c_1 \dots c_5$  - partitioning  $[0, H_{max}]$  into equally large intervals (Sec. 5) are well mapped by the decision tree to corresponding output classes, providing an accurate estimation of the local entropy values (Fig. 4).

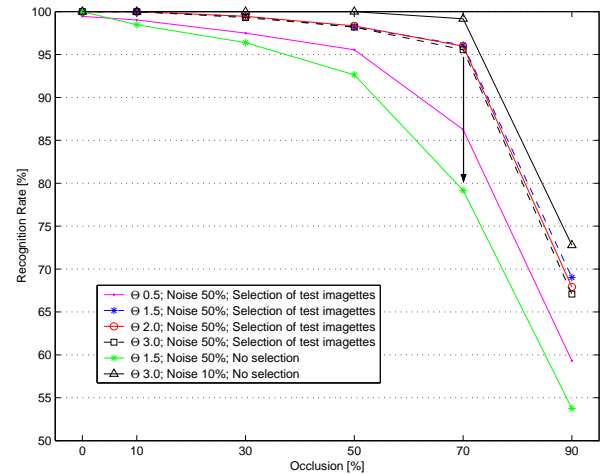
**Partial occlusion and Gaussian noise** For a thorough performance analysis in case of image corruption, we applied partial occlusion of 0 – 90% and Gaussian noise to pixel brightness. For determining the occlusion area, we selected random center positions of the black occluding squared windows within the object regions, and computed the window size related to the total number of pixels attributed to a given object. This prevents from preferring specific object regions for occlusions, and assures that the object is actually occluded according to a given occlusion rate. Fig. 5 depicts sample entropy coded color images corrupted by various degrees of occlusion. The associated histograms on imagette based object label attribution illustrate that majority voting mostly provides a both accurate and robust decision on the object identity.

**Object recognition performance** The experiments on recognition rates from occlusion and noise demonstrate the superior performance of the entropy critical method as well as the associated majority voting classifier. Fig. 6 demonstrates best performance for an interpretation of a complete image that has been treated by small Gaussian noise = 10%. However, for detection issues we have to confine to local regions such as those segmented by entropy thresholding ( $\Theta < \Theta_{max}$ ). Several models of entropy thresholding demonstrate the robustness of the system performance with respect to Gaussian noise = 50% and for varying degrees of occlusions. Note that with an entropy critical selection of 30% ( $\Theta = 1.5$ ) out of all possible test imagettes an accuracy of > 95% is achieved despite a 70% occlusion rate (blue). Considering instead *all* test imagettes for recognition (no selection), the performance would drop by more than 15% (green; arrow)! To the knowledge of the authors, the entropy critical classifier is outperforming any comparable method in the literature, since comparable local recognition results have been achieved without noise only.

**Attentive object detection** Preliminary experiments on the attentive object detection system have been performed to evaluate it in



**Figure 5.** Samples of (top down) 40%, 80% and 60% occlusion on entropy coded images and associated histogram on imagette based object classes. Majority voting provided (top down) correct/correct/incorrect identification.



**Figure 6.** Resulting recognition performance for up to 90% occlusion rates and Gaussian noise of 50% (10%).

the presence of background clutter and noise (Fig. 7). Objects from the COIL-20 database were copied into an image of an everyday environment (a) for detection. Local appearance patterns were first interpreted in a most rapid image analysis using the rapid decision tree based estimator of the entropy function (b). Applying a threshold on the estimated entropy to obtain regions of interest for further analysis ( $\Theta \leq 1.0$  was used on Fig. 7(b)). Finally, patterns in-



side the discriminative regions were mapped into eigenspace for a 1-NN analysis with respect to the sparse representation ( $\Theta \leq 1.0$ ). This results in object hypotheses for nearest neighbor distances in eigenspace  $\leq \epsilon = 0.1$ , and no labelling otherwise. The result of the interpretation (smoothed with a Gaussian filter of  $\sigma = 4.0$ ) is depicted in Fig. 7(c): most of the estimated object locations correspond with pixel coordinates of the associated object pattern in (a). E.g., attentive search for COIL object number 4 would result in the exact position of the queried object. As some artefacts may not correspond to object patterns, we expected to further disambiguate these locations from a more thorough analysis in a hierarchical framework of object recognition, dedicated to future work.

## 6 CONCLUSIONS

This work represents a thorough statistical analysis of local discriminative information for object recognition applied to images of a well cited reference database. It demonstrates that the local information content of an image with respect to object recognition provides a favourable measure to determine both sparse object models and interest operators for detection. Focusing image analysis exclusively on discriminative regions will not only result in accelerated processing but even in superior recognition performance (Sec. 5). The methods potential for applications is in object detection tasks, such as in rapid and robust video analysis.

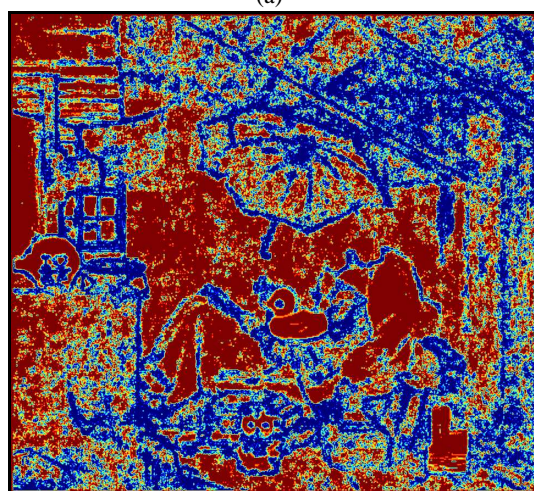
Future work will focus on finding appropriate methods to further thin out the sparse object model, and taking local topology in sub-space into concern. Furthermore, ongoing work considers to develop a grouping mechanism that would locally segment promising regions of interest for detection with the goal of cascaded object detection.

## REFERENCES

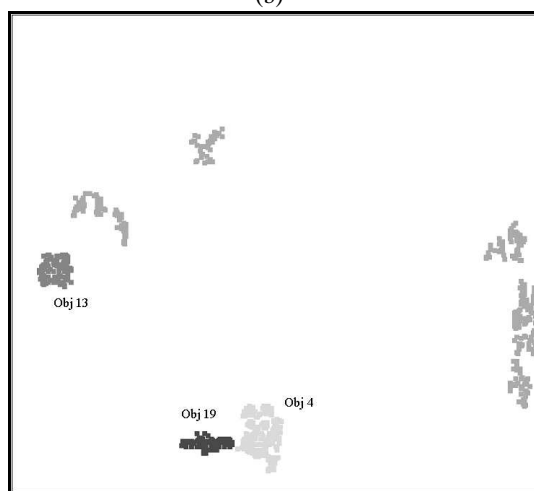
- [1] S. Agarwal and D. Roth, 'Learning a sparse representation for object detection', in *Proc. European Conference on Computer Vision*, volume 4, pp. 113–130, (2002).
- [2] V. C. de Verdière and J. L. Crowley, 'Visual recognition using local appearance', in *Proc. European Conference on Computer Vision*, (1998).
- [3] R. Fergus, P. Perona, and A. Zisserman, 'Object class recognition by unsupervised scale-invariant learning', in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 264–271, (2003).
- [4] J. H. Friedman, J. L. Bentley, and R. A. Finkel, 'An algorithm for finding best matches in logarithmic expected time', *ACM Transactions on Mathematical Software*, **3**(3), 209–226, (1977).
- [5] T. Kadir and M. Brady, 'Scale, saliency and image description', *International Journal of Computer Vision*, **45**(2), 83–105, (2001).
- [6] K. Mikolajczyk and C. Schmid, 'An affine invariant interest point detector', in *Proc. European Conference on Computer Vision*, pp. 128–142, (2002).
- [7] H. Murase and S. K. Nayar, 'Visual learning and recognition of 3-D objects from appearance', *International Journal of Computer Vision*, **14**(1), 5–24, (1995).
- [8] S. Obdrzalek and J. Matas, 'Object recognition using local affine frames on distinguished regions', in *Proc. British Machine Vision Conference*, pp. 113–122, (2002).
- [9] L. Paletta and C. Greindl, 'Context based object detection from video', in *Proc. International Conference on Computer Vision Systems*, pp. 502–512, (2003).
- [10] E. Parzen, 'On estimation of a probability density function and mode', *Annals of Mathematical Statistics*, **33**, 1065–1076, (1962).
- [11] J.R. Quinlan, *C4.5 Programs for Machine Learning*, Morgan Kaufmann, San Mateo, CA, 1993.
- [12] D.L. Swets and J. Weng, 'Using discriminant eigenfeatures for image retrieval', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **18**(8), 831–837, (1996).
- [13] M. Weber, M. Welling, and P. Perona, 'Unsupervised learning of models for recognition', in *Proc. European Conference on Computer Vision*, pp. 18–32, (2000).



(a)



(b)



(c)

**Figure 7.** Sample results of attentive object detection (Sec. 5). (a) Original frame with COIL-20 objects superimposed. (b) Corresponding entropy map using decision tree based rapid estimation (Sec. 4). (c) Detected object locations annotated by associated COIL-20 object hypothesis.