Rapid Object Recognition from Discriminative Regions of Interest[*]

Gerald Fritz, Christin Seifert, Lucas Paletta
JOANNEUM RESEARCH
Institute of Digital Image Processing
Wastiangasse 6, A-8010 Graz, Austria
{gerald.fritz,christin.seifert,lucas.paletta}@joanneum.at

Horst Bischof
Graz University of Technology
Institute for Computer Graphics and Vision
Inffeldgasse 16/II, A-8010 Graz, Austria
bischof@icg.tu-graz.ac.at

Abstract

Object recognition and detection represent a relevant component in cognitive computer vision systems, such as in robot vision, intelligent video surveillance systems, or multi-modal interfaces. Object identification from local information has recently been investigated with respect to its potential for robust recognition, e.g., in case of partial object occlusions, scale variation, noise, and background clutter in detection tasks. This work contributes to this research by a thorough analysis of the discriminative power of local appearance patterns and by proposing to exploit local information content to model object representation and recognition. We identify discriminative regions in the object views from a posterior entropy measure, and then derive object models from selected discriminative local patterns. For recognition, we determine rapid attentive search for locations of high information content from learned decision trees. The recognition system is evaluated by various degrees of partial occlusion and Gaussian image noise, resulting in highly robust recognition even in the presence of severe occlusion effects.

Introduction

Research on visual object recognition and detection has recently focused on the development of local interest operators (Kadir & Brady 2001; Mikolajczyk & Schmid 2002; Obdrzalek & Matas 2002) and the integration of local information into occlusion tolerant recognition (Weber, Welling, & Perona 2000; Agarwal & Roth 2002; Fergus, Perona, & Zisserman 2003). Recognition from local information (part based recognition) serves 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 recognition developed several viewpoints on how to advance from local to global information. (Weber, Welling, & Perona 2000) applied standard interest operators (Förstner, Harris) with the aim to determine localizable object parts for further analysis. (Weber, Welling, & Perona 2000) selected class related patterns from related fixed spatial configurations of recognizable operator responses. To avoid dependency on scale selection, (Kadir & Brady 2001; Mikolajczyk & Schmid 2002; Obdrzalek & Matas 2002) 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 (Fergus, Perona, & Zisserman 2003). While concern has been taken specifically with respect to issues of scale invariance (Lowe 1999; Kadir & Brady 2001), wide baseline stereo matching performance (Mikolajczyk & Schmid 2002; Obdrzalek & Matas 2002), or unsupervised learning of object categories (Weber, Welling, & Perona 2000; Agarwal & Roth 2002; Fergus, Perona, & Zisserman 2003), 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 of the presented work to local object recognition is, firstly, to provide a thorough analysis on the discriminative power of local appearances, and secondly, to exploit discriminative object regions to build up an efficient local appearance based representation and recognition methodology (Fig. 1). In contrast to the use of classifiers that determine discriminant features (e.g., (Swets & Weng 1996)) for recognition, our approach intends to make the actual local information content explicit for further processing. An information theoretic saliency measure is then used to construct the object model (Sec. 2) or to determine discriminative regions of interest for detection and recognition purposes (Sec. 3). In this work we focus on the investigation of the local information content, and how this leads to an early mapping to reject meaningless image regions. While this study primarily considers invariance to rotation of the object around the vertical axis as well as to translation (Sec. 5), we think that an extension to scale and (horizontal) rotation invariant local features (such as reported in (Lowe 1999) or (Mikolajczyk & Schmid 2002)) would be straight forward, by taking the discriminative potential of the features into account.

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 sec-

[*]This work is funded by the European Commission's project DETECT under grant number IST-2001-32157, and by the Austrian Joint Research Project Cognitive Vision under sub-projects S9103-N04 and S9104-N04. This work has been carried out within the K plus Competence Center ADVANCED COMPUTER VISION.

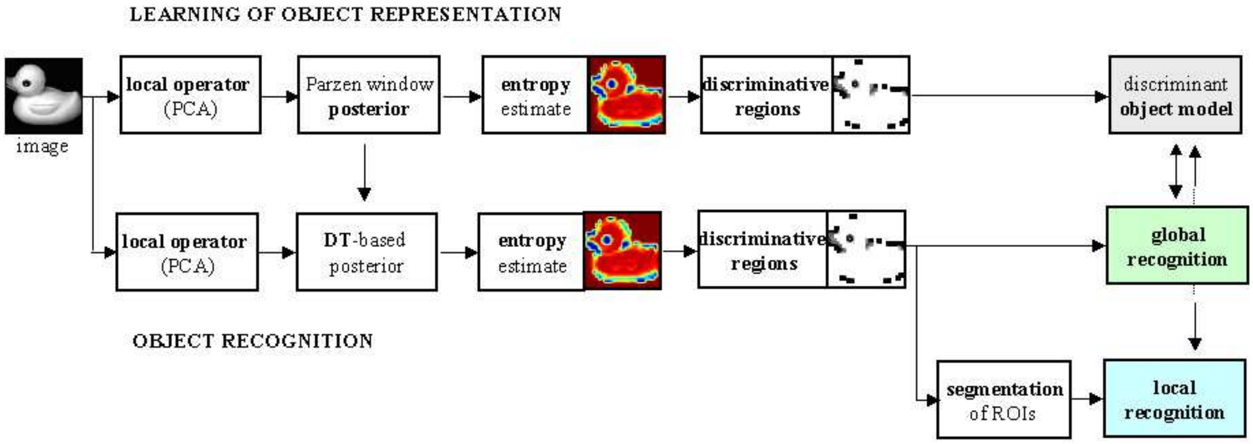

Figure 1: Concept of the entropy based recognition model. (Top) From a subspace feature selection (e.g., PCA) we compute the posteriors and the local information content. Responses of discriminative regions with entropy values below a predefined threshold are stored in the object model. (Bottom) For 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.

ond 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).

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*[1] 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.

Note that discriminative regions correspond here to regions in the image that contain information with respect to object recognition. Feature extraction is here applied from principal component analysis and followed by an estimate of the local posterior to derive the local information content. In contrast to comparable work on the extraction of discriminative basis libraries (e.g., (Saito *et al.* 2002)) for an efficient construction of feature subspaces, we specifically exploit the local information content value to derive efficient thresholds for pattern rejection, with the purpose to determine both representation and region of interest.

Local distributions in subspace

We use a principal component analysis (PCA, (Murase & Nayar 1995)) 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 $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_i, \ldots, \mathbf{x}_N\}$, 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 $P(o_k|\mathbf{g}_i)$, $k = 1 \ldots \Omega$, Ω is the number of instantiations of the object class variable O. The Shannon entropy denotes

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

We approximate the posteriors at $\mathbf{g}_i$ using only samples $\mathbf{g}_j$ inside a Parzen window (Parzen 1962) of a local neighborhood ϵ, $||\mathbf{g}_i - \mathbf{g}_j|| \leq \epsilon$, $j = 1 \ldots J$. We weight the contributions of specific samples $\mathbf{g}_{j,k}$ - labeled 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 favor 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 $\mathbf{g}_i$.

Discriminative object regions

It is obvious that the size of the local ϵ-neighborhood will impact the distribution and thereby the recognition accuracy (Fig. 3, Sec. 5), highly depending on the topology of object related manifolds in subspace. One can construct

[1] imagettes denote subimages of an object view (de Verdiére & Crowley 1998)

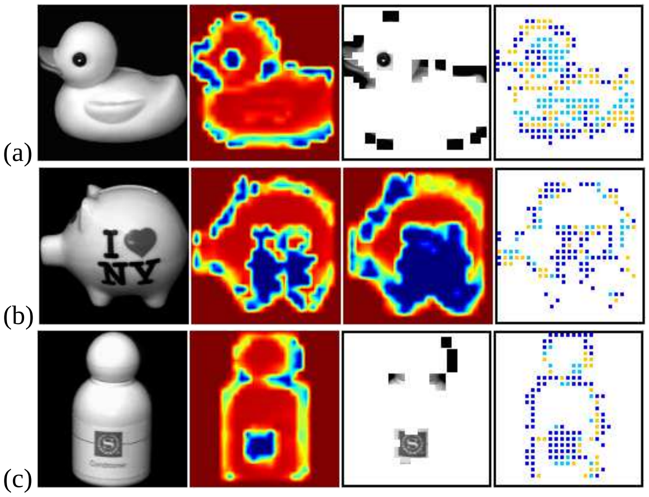

Figure 2: Sample COIL-20 objects (a) o_1, (b) o_{13}, (c) o_{16} at view 0; each with - left to right - (i) original frame, (ii) entropy image (from 9x9 pixel imagettes; entropy coded color palette from low=blue to high=red), (iii) local appearances with $\Theta \leq 0.5$ in a,c (entropy image from 15x15 pixel imagettes in b), and (iv) accuracy-coded images (accuracy blue=true, white=false).

an entropy-coded image of an object view from a mapping of local appearances $\mathbf{x}_i$ to corresponding entropy estimates (Fig. 2 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 9×9 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 (de Verdiére & Crowley 1998), in 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 (de Verdiére & Crowley 1998), or from selected, cluster specific prototype points (Weber, Welling, & Perona 2000; Fergus, Perona, & Zisserman 2003) 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

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

A specific choice on the threshold Θ consequently determines both storage requirements and recognition accuracy (Sec. 5). The question on how to practically select Θ must be answered in response to application related considerations, such as, requested recognition accuracy, etc. One recommended strategy to determine Θ would be to start with the Θ_{max}, thereby requesting a maximum of resources, and then to gradually reduce Θ so that a critical limit will be empirically encountered.

To speed up the matching we use efficient memory indexing of nearest neighbor candidates described by the adaptive *K-d* tree method (Friedman, Bentley, & Finkel 1977).

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 ϵ-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 analysed 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 (Paletta & Greindl 2003). The entropy threshold Θ 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 ϵ-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.

Discriminative regions from decision trees

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 segmentation function would work in terms of a point of interest operator (POI) but be tuned from a discriminative objective function, i.e., 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 (Quinlan 1993) 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 (Quinlan 1993).

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.

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 (Murase & Nayar 1995).

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 128×128 pixels. Analysis was performed with various imagette sizes: 9×9 (selected for further experiments), 15×15, and 21×21 pixels, resulting in larger discriminative regions (normalised per entropy maximum; e.g., in Fig. 2b) 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 ϵ to determine the training samples in a local neighborhood (Sec. 2.1) that will be considered for the posterior estimation. Fig. 3 shows recognition rates (using a MAP classifier and images degraded with 10% Gaussian noise, operated on all *single* imagettes) with various values for neighborhood ϵ and entropy threshold Θ. 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.

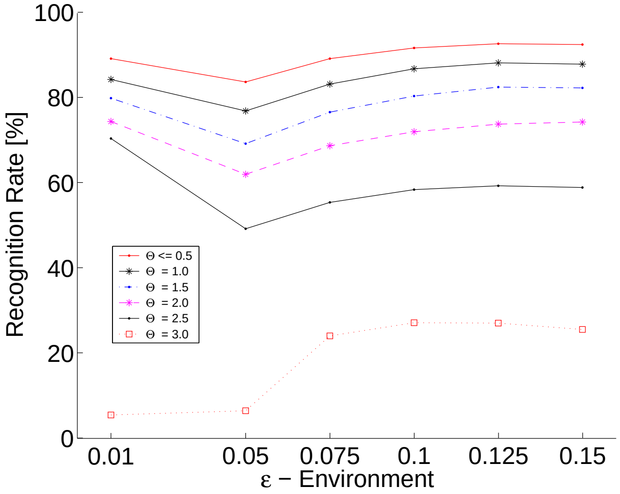

Figure 3: Recognition performance using MAP classification on samples of a neighborhood ϵ. Rejecting imagettes with entropy $\hat{H}(O|\mathbf{g}_i) > \Theta$, $\Theta = 2.0$, may dramatically increase accuracy of overall object recognition.

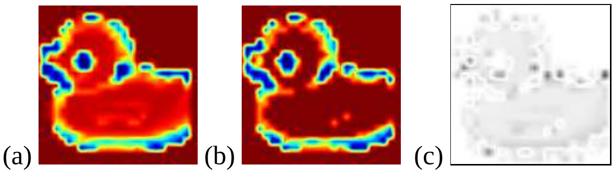

Figure 4: Estimation of entropy from decision trees. (a) Entropy estimation from Parzen windows I_P (Sec. 2.1, color code see Fig. 2), (b) entropy estimation from decision tree I_T (Sec. 4), (c) difference image $I_D = I_P - I_T$ (grayvalue coded for $[0, H_{max}]$ from white to black).

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 ≈ 25 attribute queries. A difference image (Fig. 4c) reflects the negligible errors and confirms that discriminative regions can be both reliably and rapidly estimated from C4.5 decision trees.

maps ↦	$[0, H_{c_1})$	$[H_{c_1}, H_{c_2})$	$[H_{c_2}, H_{c_3})$	$[H_{c_3}, H_{c_4})$	$[H_{c_4}, H_{max}]$
$[0, H_{c_1})$	38146	5546	2981	1425	218
$[H_{c_1}, H_{c_2})$	1820	44418	1382	624	72
$[H_{c_2}, H_{c_3})$	1495	2094	41989	2548	190
$[H_{c_3}, H_{c_4})$	1036	1136	3725	40775	1644
$[H_{c_4}, H_{max}]$	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 $c_1 \cdots c_5$ - 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.

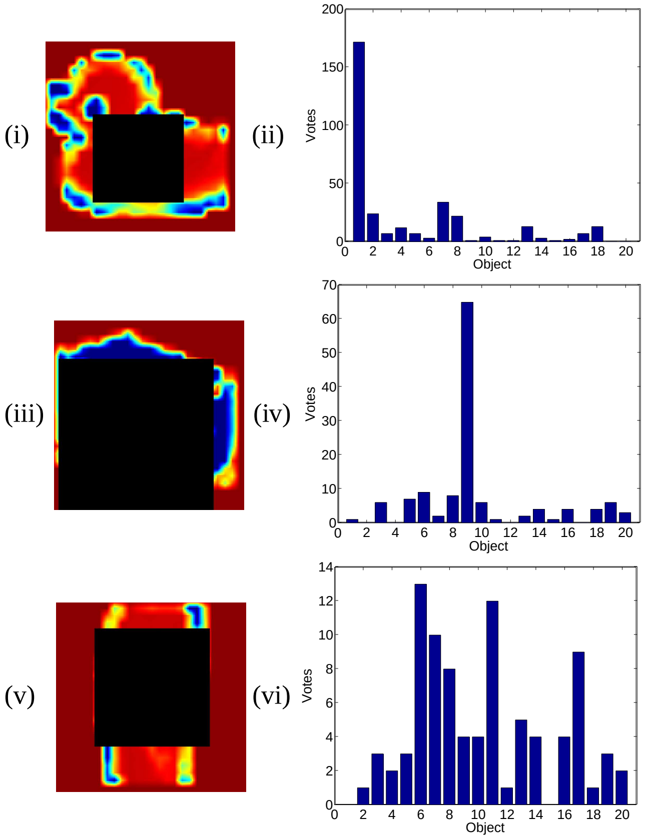

Figure 5: Samples of (top down) 40%, 80% and 60% occlusion on entropy coded images and associated histogram on imagette based object classes. Majority voting provided (top down) correct/correct/incorr. identification.

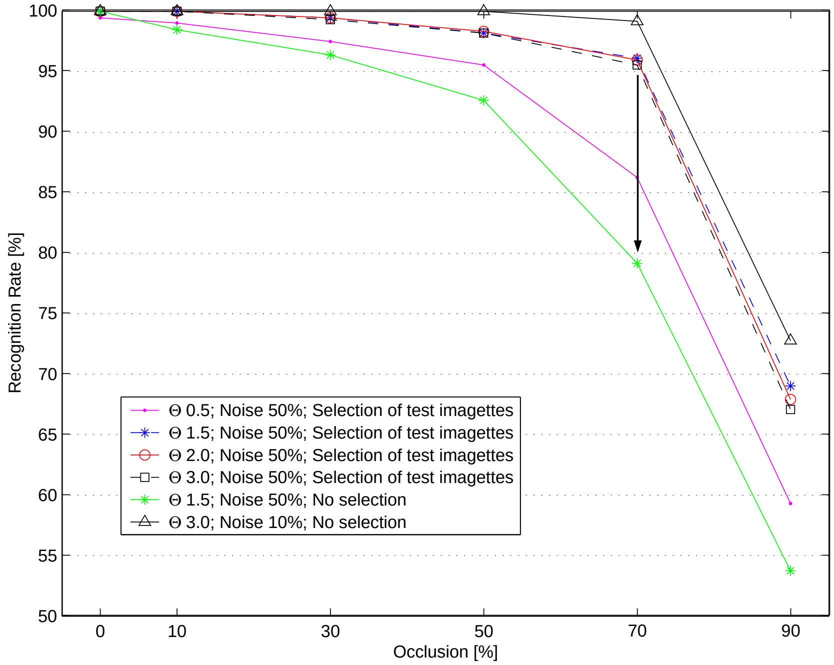

Figure 6: Resulting recognition performance for up to 90% occlusion rates and Gaussian noise of 50% (10%).

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 favorable 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 subspace 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

Agarwal, S., and Roth, D. 2002. Learning a sparse representation for object detection. In *Proc. European Conference on Computer Vision*, volume 4, 113–130.

de Verdiére, V. C., and Crowley, J. L. 1998. Visual recognition using local appearance. In *Proc. European Conference on Computer Vision*.

Fergus, R.; Perona, P.; and Zisserman, A. 2003. Object class recognition by unsupervised scale-invariant learning. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 264–271.

Friedman, J. H.; Bentley, J. L.; and Finkel, R. A. 1977. An algorithm for finding best matches in logarithmic expected time. *ACM Transactions on Mathematical Software* 3(3):209–226.

Kadir, T., and Brady, M. 2001. Scale, saliency and image description. *International Journal of Computer Vision* 45(2):83–105.

Lowe, D. 1999. Object recognition from local scale-invariant features. In *Proc. International Conference on Computer Vision*, 1150–1157.

Mikolajczyk, K., and Schmid, C. 2002. An affine invariant interest point detector. In *Proc. European Conference on Computer Vision*, 128–142.

Murase, H., and Nayar, S. K. 1995. Visual learning and recognition of 3-D objects from appearance. *International Journal of Computer Vision* 14(1):5–24.

Obdrzalek, S., and Matas, J. 2002. Object recognition using local affine frames on distinguished regions. In *Proc. British Machine Vision Conference*, 113–122.

Paletta, L., and Greindl, C. 2003. Context based object detection from video. In *Proc. International Conference on Computer Vision Systems*, 502–512.

Parzen, E. 1962. On estimation of a probability density function and mode. *Annals of Mathematical Statistics* 33:1065–1076.

Quinlan, J. 1993. *C4.5 Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann.

Saito, N.; R.R.Coifman; Geshwind, F.; and Warner, F. 2002. Discriminant feature extraction using empirical probability density estimation and a local basis library. volume 35, 2841–2852.

Swets, D., and Weng, J. 1996. Using discriminant eigenfeatures for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 18(8):831–837.

Weber, M.; Welling, M.; and Perona, P. 2000. Unsupervised learning of models for recognition. In *Proc. European Conference on Computer Vision*, 18–32.