Entropy based Saliency Maps for Object Recognition[1]

Gerald Fritz[⋆], Christin Seifert[⋆], Lucas Paletta[⋆] and Horst Bischof[†]

[⋆]Joanneum Research
Institute of Digital Image Processing
Wastiangasse 6, A-8010 Graz, Austria
E-mail: {gerald.fritz, christin.seifert, lucas.paletta}@joanneum.at

[†]Graz University of Technology
Institute for Computer Graphics and Vision
Inffeldgasse 16/II, A-8010 Graz, Austria
E-mail: bischof@icg.tu-graz.ac.at

Abstract

Object identification from local information has recently been investigated with respect to its potential for integration and robust recognition. In contrast to existing approaches, we do not use generic interest operators but select regions of interest from top-down information, i.e., with respect to object recognition. Discriminative regions are determined from the information content in the local appearance patterns (imagettes) and consequently enable to model sparse object representation and attention based recognition using decision trees. Recognition performance from single imagettes 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.

1 Introduction

Cognitive computer vision systems must address 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.

Research on visual object recognition and detection has recently focused on the development of generic local interest operators and the class based integration of local information into occlusion tolerant recognition. (Weber et al., 2000) applied standard interest operators (Förstner, Harris) with the aim to determine localizeable object parts for further analysis. To avoid dependency on scale selection, (Kadir and Brady, 2001; Mikolajczyk and Schmid, 2002; Obdrzalek and Matas, 2002) introduced interest point

[1] This work is funded by the European Commission's IST project DETECT under grant number IST-2001-32157, and the Austrian Joint Research Project 'Cognitive Vision', sub-projects S9103-N04 and S9104-N04.

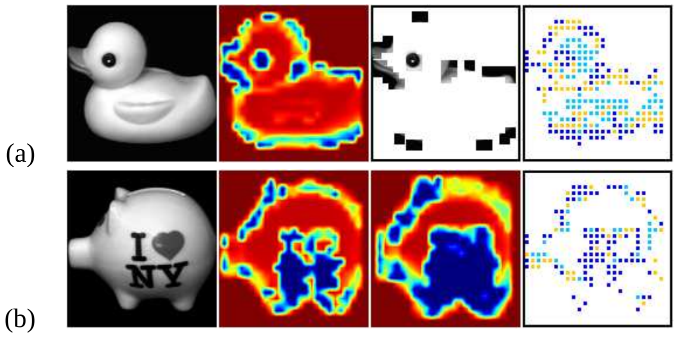

Figure 1: Sample COIL-20 objects (a) o_1, (b) o_{13} with - left to right - (i) original frame, (ii) entropy saliency map (from 9x9 pixel imagettes; entropy from low=blue to high=red), (iii) local appearances with $\Theta \leq 0.5$ in a) and entropy image from 15x15 pixel imagettes in b), and (iv) accuracy-coded images (accuracy blue=true, white=false).

detectors that derive scale invariance from local scale saliency. These operators proved to further improve recognition from local photometric patterns (Fergus et al., 2003).

The key contribution of the presented work is to investigate top-down influence on interest operators on the basis of an information theoretic approach. Entropy based saliency provides discriminative object regions that are shown to build up a highly efficient and *sparse object representation*. This saliency measure enables *attentive recognition* from rapid local entropy estimation derived from inductive inference for decision tree based mapping. The method is evaluated on images degraded with Gaussian noise and different degrees of partial occlusions using the COIL database (Sec. 4).

2 Entropy-based salience

Local regions in the object views that are both discriminative and robustly indicate the correct object label provide the reference imagettes[2] for the object representation. We use a principal component analysis (PCA, (Murase and Nayar, 1995)) calculated on local image windows of size $w \times w$ to form the basis for our local low dimensional representation. 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 $P(o_k|\mathbf{g}_i)$, $k = 1 \dots \Omega$, 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 of a local neighborhood ϵ, $||\mathbf{g}_i - \mathbf{g}_j|| \leq \epsilon$. We weight the contributions of specific

[2] imagettes denote subimages of an object view (de Verdiére and Crowley, 1998)

samples $\mathbf{g}_{j,k}$ that should increase the posterior estimate $P(o_k|\mathbf{g}_i)$ by a Gaussian kernel function value $\mathcal{N}(\mu = \mathbf{g}_i, \sigma = \epsilon/2)$ in order to favour samples with smaller distance to observation $\mathbf{g}_i$. 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 $\mathbf{g}_i$. It is obvious that the size of the local ϵ-neighborhood will impact the distribution and thereby the recognition accuracy (Fig. 2, Sec. 4). One can construct an entropy based saliency map of an object view from a mapping of local appearances $\mathbf{g}_i$ to corresponding entropy estimates (Fig. 1).

From discriminative regions we proceed to *entropy thresholded* in contrast to extensive (de Verdiére and Crowley, 1998) object *representations*. The proposed object model includes only *selected* reference points for nearest neighbor classification, storing exclusively those $\mathbf{g}_i$ with $\hat{H}(O|\mathbf{g}_i) \leq \Theta$. A specific choice on the threshold Θ consequently determines both storage requirements and recognition accuracy (Sec. 4).

3 Recognition from local information

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

1. **Mapping** of imagette patterns into eigenspace.
2. **Probabilistic interpretation** to determine entropy $\hat{H}(O|\mathbf{g}_i)$. For rapid interpretation, decision trees provide sufficiently well approximations.
3. **Rejection** of imagettes with ambiguous information that might degrade accumulating evidence for a correct object hypothesis (Paletta and Greindl, 2003).
4. **Nearest neighbor classification** of selected imagettes within ϵ-environment.
5. **Majority voting** for object identifications over a full image nearest neighbor analysis. 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.

Inductive inference on the appearance patterns with respect to discriminative entropy intervals provides a most rapid decision tree mapping (Quinlan, 1993) (Fig. 2b,c) to focus attention on most salient object regions.

4 Experiments and conclusion

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

Single imagette interpretation Experiments were applied on 72 (test: 36) views from 20 objects of the COIL-20 database (Murase and Nayar, 1995). Analysis was performed with 9×9 pixel imagettes mapped onto a 20-dimensional eigenspace. Fig. 2

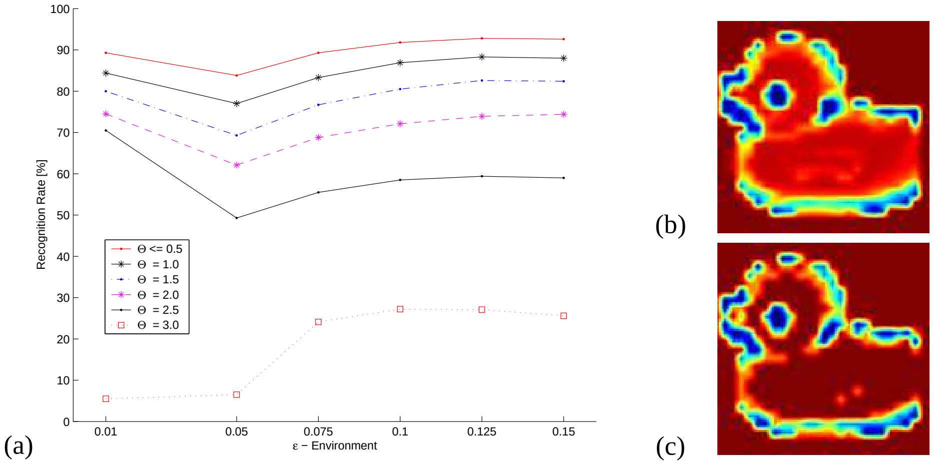

Figure 2: (a) Recognition performance using MAP classification on samples of a neighborhood ϵ. Rejecting imagettes with entropy $\hat{H}(O|\mathbf{g}_i) > \Theta$, $\Theta = 2.0$, may dramatically increase accuracy of overall object recognition. (b) Entropy estimation from Parzen window analysis, (c) rapid decision tree based estimation.

shows higher recognition rates (using a MAP classifier on 10% Gaussian noise degraded images) for discriminative ($\Theta < \Theta_{max}$) single imagettes (selecting $\epsilon = 0.1$ for further processing).

Partial occlusions Fig. 3 depicts a sample entropy coded image corrupted by occlusion. The associated histogram on imagette based object label attribution illustrates that majority voting mostly provides a both accurate and robust decision on the object identity. 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. 3b) demonstrates the robustness of the performance with Gaussian noise = 50% 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% (arrow).

Conclusion This work represents a 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 a sparse object model, to accelerate processing and to provide superior recognition performance even from degraded image content. The methods potential for applications is in object detection tasks, such as in rapid and robust video analysis.

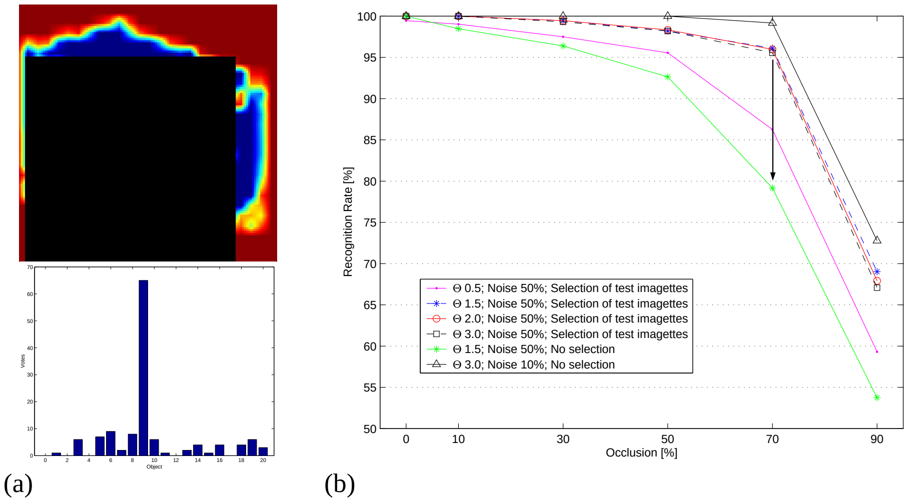

Figure 3: (a) 80% occlusion on entropy coded images and associated object class histogram. (b) Recognition performance for different occlusion rates and Gaussian noise.

References

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*, pages 264–271.

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

Mikolajczyk, K. and Schmid, C. (2002). An affine invariant interest point detector. In *Proc. European Conference on Computer Vision*, pages 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*, pages 113–122.

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

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

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