ESANN'2003 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 23-25 April 2003, d-side publi., ISBN 2-930307-03-X, pp. 503-508

Semi-automatic acquisition and labelling of image data using SOMs

Gunther Heidemann, Axel Saalbach, Helge Ritter

Bielefeld University, Neuroinformatics Group,
P.O. Box 100131, D-33501 Bielefeld, Germany

Abstract. Application of neural networks for real world object recognition suffers from the need to acquire large quantities of labelled image data. We propose a solution that acquires images from a domain at random and structures the data in two steps: Data driven mechanisms extract windows of interest, which are clustered by a SOM. Regions of the SOM in which objects form clusters serve as "suggestions" for categories. An interactive visualisation of the SOM combined with distance measures allows the user to determine classes and build training sets. By this means, large labelled data sets for a neural classifier can be easily generated.

1 Introduction

Recognition of objects and other visual entities is one of the major problems in computer vision, but still the bridge from the pixel appearance of objects to the semantic level can be built only in special cases. Where objects of interest are physically available, *appearance based* representations have become popular. We will shortly outline this idea, its major problem, and a possible solution.

1.1 Appearance based recognition

Object knowledge can be acquired from mainly two sources: Sample images or human expertise. But designing e.g. geometrical object models from human knowledge is costly, moreover, such models are difficult to match with the signal. Therefore, designing explicit representations is limited to special cases.

The appearance based approach memorizes features of sample views instead of models. By this means, the pixel appearance has not to be modelled explicitly in all variations caused by varying pose, lighting or occlusions. Such methods have been applied successfully in object- [7] or face recognition [6].

1.2 Problems of the appearance based approach

The benefit of the appearance based approach – memorizing samples – is also its Achilles heel. Even for simple, rigid objects it is hard to acquire images under all possible real world conditions, because the product space of poses, lighting conditions etc. would have to be sampled. Solutions to this problem can be sought in mainly three directions:

(1) *Smart solutions:* The best solution would be to separate all types of appearance variations, e.g. pose from lighting, using a factorial code, and to derive this code from samples of the factors only. So when a recognition system has seen (*i*) an object in a limited set of poses and (*ii*) relevant basic lighting conditions, it should recognise the object in arbitrary pose under arbitrary lighting. Unfortunately, such methods are not yet generally available.

(2) *Brute force solutions:* It is possible that no smart solutions exist. The human visual system is trained over years by an enormous amount of visual stimuli, so possibly nature makes use of a “brute force” strategy. As computer capacity increases, the question arises how computers could do the same.

(3) *A factorial solution acquired from non-factorial sampling:* Currently, the most hopeful approach appears to be a compromise of the other two: Building a “smart representation from brute force sampling”. The idea is that a factorial solution can be achieved only *after* a great number of samples has been memorized and further processed. Since neural networks have in principle the ability to extract factorial representations, it makes sense to investigate how such large labelled image data sets could be produced.

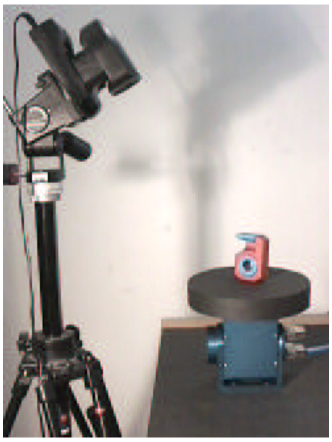
2 System description

The problem that prevents the use of huge quantities of visual data in machine vision is not primarily lack of memory but *acquisition of labelled data.* In [7] objects are presented in isolation on a turntable, which allows to acquire automatically many labelled images. However, this method cannot be extended to real world complexity. This limitation could be overcome only if

- images of *real scenes* can be used that show *many* objects at a time in *unknown poses*,
- the objects can be *extracted automatically* and
- *no hand-labelling* of single objects is required.

The ultimate goal would be a system that acquires image data by itself from the world (e.g. using a mobile platform), extracts the objects and forms categories by itself. So the human designer would be needed only to label objects *after* the categories were found automatically.

As a first step we propose a semi-automatic system for image acquisition, labelling and recognition of objects or other visual entities. It uses arbitrary images as a starting point from which data driven attentional mechanisms extract



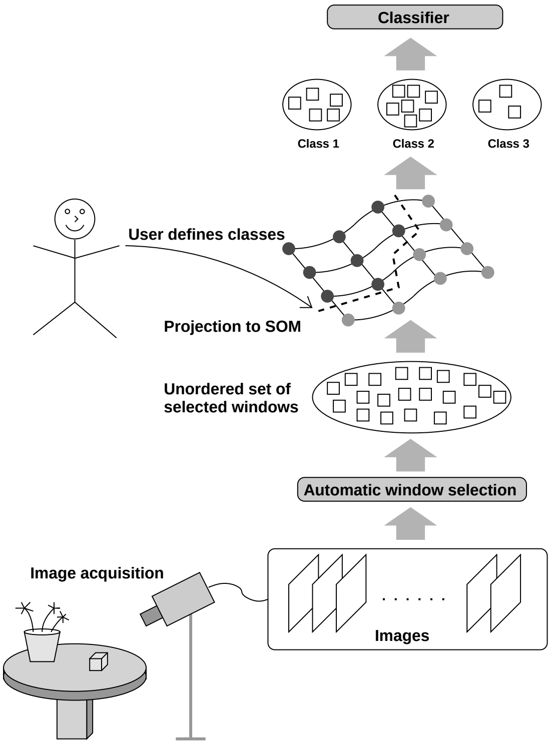


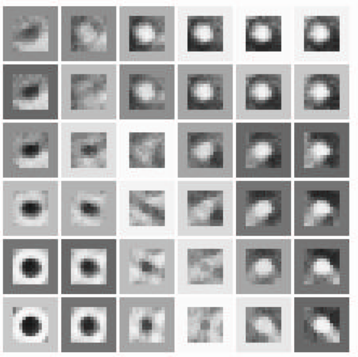
Figure 1: Above: Turn-tilt table for object presentation with active camera. Right: Semi-automatic data acquisition. Automatically selected windows are clustered using a SOM, then the user selects regions of nodes as classes to train a classifier.

windows of interest. From this data distribution a self-organizing map (SOM) [4] is trained which partitions the data into clusters. A graphical user interface allows an easy labelling of the data to obtain training sets for a classifier.

2.1 Acquisition of unordered image windows

As a “first approximation” to real world complexity, we use a motorised turn-tilt table with two degrees of freedom, on which several objects can be fastened. An active camera follows the movements of the turn-tilt table and takes pictures (Fig. 1). Objects of two entirely different domains are used: A flowerpot with blossoms and wooden toy pieces (Baufix) which can be assembled to complex objects (Fig. 2). So we don’t follow the scheme of single object acquisition which would lead to images associated with classes, but simulate the situation of the real world: The camera gets data from different domains all at once. Moreover, the joint angles of the turn-tilt table are not memorized, so object pose information is not available.

The images are not memorized but only windows for which “interestingness” can be detected on the signal level. Three different saliency features are currently used: (*i*) Edges and corners detected by the operator of Harris and Stephens [1], local symmetry as proposed by Reisfeld et al. [9] and the



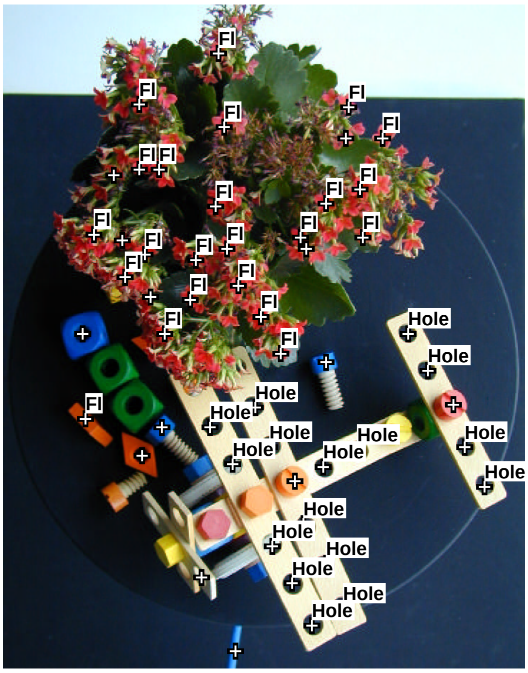


Figure 2: Above: 6 × 6-SOM trained on windows sampled from the "flower + Baufix" setup. The SOM is divided in areas corresponding to holes (of Baufix) and blossoms, the rest is unclear. Right: Interest points and classification results.

centers of salient homogeneous colour regions using the algorithm proposed by Rehrmann and Priese [8]. Fig. 2 shows for simplicity only the interest points derived from symmetry, around each a window is cut out.

2.2 Semi-automatic structuring and labelling

We use the system VALT (Visualization And Labeling Toolkit) [10] to divide the unordered windows into classes. Each window has the same size $w \cdot h$ and is regarded as a vector $\vec{x} \in \mathbb{R}^d$ with dimension $d = 3 \cdot w \cdot h$ (for three colour channels). VALT trains a SOM [4] using all windows $\vec{x}$. The nodes of the SOM can be visualised either using their weight vectors in pixel space or their best match example. The SOM then shows regions which correspond to visual entities found by the window detection module. The user can decide which nodes form relevant visual classes and join them graphically by drawing a line around them (mouse input). When a class was formed, a numerical identifier and a label have to be associated. All samples which are best match to the selected nodes then belong to their assigned class. Nodes which are not of interest or for which membership is unclear may be put into a rejection class. To facilitate decisions, several measures for distances between the nodes like the *Unified Distance Matrix* [11] or the *Kohonen Projection Method* [5] can be displayed as coloured background to the SOM.

The labelled samples are to be used as a training set for a suitable classifier, here we apply the VPL-architecture described e.g. in [3]. This system combines a trainable feature extraction based on local PCA [6] with several neural classifiers. After the training phase, the classifier assigns classes to the extracted interest points close to real time.

3 Results

Fig. 2 shows a scene on the turn-tilt table with a flowerpot and Baufix toy pieces. To keep the demonstration simple, only symmetry based interest points after [9] are displayed which tend to be centered at holes and on blossoms. The used window has approximately the size of a hole. For simplicity, only a small 6×6–SOM is used, so only the best represented classes form visible regions, here holes of Baufix-bars and blossoms. Nodes belonging to these two classes can be recognized, the other nodes remain unclear and are assigned to the rejection class. Fig. 2 shows the results of a trained VPL-classifier [3]. Label "Fl" stands for Flower, unlabelled interest points belong to the rejection class.

In more difficult experiments which can not be described in detail all three attentional mechanisms and SOMs up to size 20×20 were used. In this case up to 16 classes could be formed, e.g. for blossoms, dry blossoms, tips of green leaves as well as holes and bolt heads of different types. However, also unexpected categories turned up like the space between the bars of the toy airplane. Collecting 10000 windows takes only about three minutes on the turn-tilt table and approximately the same time is needed to train the SOM. If an experienced user draws the class borders, the entire training set can be designed in about ten minutes.

For comparison, SOMs were trained also on the Columbia Object Image Library [7]. Despite its size, this image collection proves to be of less complexity than the current setup. In another application the same system (without the turn-tilt table) was applied for recognition of pointing gestures [2], where acquisition of "labelled" images showing definite pointing directions is particularly tiresome. Using VALT, it is sufficient to take images of subjects just pointing somewhere and to assign pointing directions later for many images in common.

4 Discussion and outlook

We have presented a system that facilitates rapid acquisition and labelling of image data from unstructured images, using a SOM for clustering and visualisation. By this means the system makes "suggestions" which data might be put together in a class. The considerably reduced effort for acquisition of labelled data opens up the possibility to exploit much better the primal advantage of neural networks — to learn from samples.

What is the "cost" of our approach? (1) Recognition is limited to entities that can be extracted by domain independent methods with sufficient stability. However, only such entities that *are* stable become visible in the SOM, others appear as clutter, so at least those objects assigned to classes are sure to be stable. (2) Categorisation can only in part be decided by the user. Clusters in the SOM are suggestions for classes derived from low level proximity. This prohibits arbitrary class formation, however, this restriction also ensures that no classes are formed which can't be recognized on the low level.

Since window selection plays a key role, future work will be aimed to stabilisation of the interest point detectors. Another direction of development is to include more features into the SOM, which works by now on the pixel level. Since eigenspace projections did not yield substantial improvements, shape independent features like local colour histograms appear to be promising.

References

[1] C. Harris and M. Stephens. A Combined Corner and Edge Detector. In *Proc. 4th Alvey Vision Conf.*, pages 147–151, 1988.

[2] G. Heidemann, R. Rae, H. Bekel, I. Bax, and H. Ritter. Integrating context-free and context-dependent attentional mechanisms for gestural object reference. In *ICVS 03*, Graz, Austria, 2003. Accepted.

[3] G. Heidemann and H. Ritter. Combining Gestural and Contact Information for Visual Guidance of Multi-Finger Grasps. In M. Verleysen, editor, *Proc. ESANN 02*, Bruges, Belgium, 2002. d-side publications.

[4] Teuvo Kohonen. *Self-Organizing maps.* Springer series in information science. Springer, Berlin, Heidelberg, New York, third edition, 2001.

[5] M. A. Kraaijfeld, J. Mao, and A. K. Jain. A Nonlinear Projection Method Based on Kohonen's Topology Preserving Maps. *IEEE. Trans. on Neural Networks*, 6(3):548–559, 1995.

[6] B. Moghaddam and A. Pentland. Probabilistic Visual Learning for Object Representation. *IEEE TPAMI*, 19(7):696–710, 1997.

[7] H. Murase and S. K. Nayar. Visual Learning and Recognition of 3-D Objects from Appearance. *Int'l J. of Computer Vision*, 14:5–24, 1995.

[8] V. Rehrmann and L. Priese. Fast and Robust Segmentation of Natural Color Scenes. In *Proc. 3rd Asian Conf. Comp. Vision*, Hongkong, 1998.

[9] D. Reisfeld, H. Wolfson, and Y. Yeshurun. Context-Free Attentional Operators: The Generalized Symmetry Transform. *Int'l J. of Computer Vision*, 14:119–130, 1995.

[10] A. Saalbach. Self-Organizing Maps zur halbautomatischen Erzeugung datennaher Klasseneinteilungen. Master's thesis, Univ. Bielefeld, 2001.

[11] A. Ultsch and H. P. Siemon. Kohonen's self organizing feature maps for exploratory data analysis. In *Proc. ICNN'90*, Netherlands, 1990.