A multi-purpose visual classification system

Gunther Heidemann

Universität Bielefeld, AG Neuroinformatik, Germany
gh@techfak.uni-bielefeld.de

Abstract. A computer vision system which can be trained to classification tasks from sample views is presented. It consists of several artificial neural networks which realize local PCA with subsequent expert nets as classifiers. The major benefit of the approach is that entirely different tasks can be solved with one and the same system without modifications or extensive parameter tuning. Therefore, the architecture is an example for the potential which lies in view based recognition: Making complicated tasks solvable with less and less expert knowledge.

1 Introduction

Artificial Neural Networks (ANN) have the advantage that they can acquire knowledge from examples. Therefore they are an ideal tool in image processing dealing with data that cannot easily be captured by explicit geometrical modeling. However, ANN cannot be applied directly to the raw pixel data because the “curse of dimensionality” would require nets of enormous size, and, even worse, huge sets of training images. Hence, a feature extraction stage that captures the interesting part of the image variance but removes redundancy has to precede classification by ANN. The problem is, that once features have to be “designed” by a human expert, the advantages of ANN evaporate. Consequently, an ANN-based visual classification system has to provide both an adaptive feature extraction and a trainable feature classifier.

A well known solution to reduce dimensionality and exploit redundancy to stabilize the detected features is principal component analysis (PCA), as outlined e.g. in [18]. In most cases the image data are projected to the \( N_P \) principal components (PCs) with the largest eigenvalues, where \( N_P \) is much smaller than the original dimensionality (i.e. the number of pixels). Moreover, interpolation between different object views becomes much simpler in the projection space than in the original pixel space [16]. Being a linear method, the limitation of PCA is that it can capture only the “global” variance of the data, which is not necessarily the most relevant data variation with respect to a given computer vision task. A suitable non-linear extension of PCA is local PCA[22], which partitions the data into subsets for which PCA performs well. Because of the advantages of local over normal PCA, it is a consequent next step to look for a classification system which can exploit these advantages.

In this contribution, a three-stage system incorporating several ANN of different types is presented [6]. The first two stages realize a feature extraction
by local principal component analysis (PCA), the last classifies the feature vectors using expert ANNs. This architecture involves a much greater variety of highly specific filters than normal PCA. This approach can deal with very high dimensional input (e.g. a whole image with three colour channels) as well as small image patches which are derived from prior processing stages. Hence, the system can be applied to a variety of visual classification tasks ranging from object recognition and pose estimation to region segmentation. Since the system is trained from sample views – instead of explicitly geometry modeling – it can be counted among the “view based” approaches. The system has been used stand alone in robotic tasks [9] and as the component of a larger hybrid system where it provides classification hypotheses to a semantic network [2].

Fig. 1. The neural VPL-classification system maps an image patch or a complete image \( x \) to a vector valued output \( y \) in three steps: Left, the best match reference vector is determined. Middle: \( x \) is projected to the local PCs which were calculated by the attached PCA-net. Right: The projection is classified by the expert LLM-net.

2 The neural classification system

The trainable classifier performs a mapping \( x \rightarrow y, x \in \mathbb{R}^M, y \in \mathbb{R}^N \) [6]. In computer vision, \( x \) consists of the pixels of an appropriately sized window of an image (grey level or colour). The window vector \( x \) can be mapped to a continuous valued output vector, e.g. for the estimation of pose parameters (angles). For classification tasks, the output is a (discrete valued) class \( k \). In this case, one separate output channel for each of the \( N \) classes is used. Training is performed with labeled sample windows \( x_i^{Tr} \) and binary output vectors \( (y_i^{Tr})_j = \delta_{ij}, i, j = 1 \ldots N \) coding class \( i \). Classification of unknown windows \( x \) is carried out taking the class \( k \) of the output component with maximal value:

\[
k = \arg \max_j ((y(x))_j).
\]

The classifier combines visual feature extraction and classification in a three-stage architecture called VPL, which stands Vector quantization, PCA and
LLM-network. The vector quantization is carried out on the raw image windows to provide a first data partitioning using $N_{VQ}$ reference vectors $r_i \in \mathbb{R}^M, i = 1 \ldots N_{VQ}$. A suitable algorithm proved to be activity equalization vector quantization (AEV) [10] since one of the major difficulties in high dimensional spaces is codeword under-utilization [5]. Generation of a homogeneously used codebook using incremental algorithms like competitive learning (e.g. “winner takes all”) is difficult because the huge volume of a high dimensional space usually cannot be “filled” by the provided data. In this case, only few reference vectors are attracted to the data whereas other reference vectors remain outliers.

Many solutions of this problem have been proposed like the neural gas [14] or the Kohonen self-organizing map [11,13]. However, since these methods introduce interactions between the reference vectors, parameters have to be chosen appropriately which turns out to be difficult in cases where the designer has no intuition. Algorithms which are designed specially to provide a homogeneous codebook utilization are e.g. AEV [10] or frequency sensitive competitive learning [1,3,4] since they take into account the codeword access frequencies.

In the VPL-architecture, to each reference vector $r_i$ of the primary vector quantization a single layer feed forward network is attached for the successive calculation of the principal components (PCs). The PCA-nets are trained as proposed by Sanger [20]. Currently, the number of PCs $N_{PCA}$ is fixed for all reference vectors, in future extensions these numbers will be determined from the local eigenvalue spectra. The input $x$ is projected to the $N_{PCA}$ PCs with the largest eigenvalues: $x \rightarrow p_l(x) \in \mathbb{R}^{N_{PCA}}, l = 1 \ldots N_{VQ}$. $p_l(x)$ can be regarded as feature vector of input $x$.

To each of the $N_{VQ}$ PCA-nets one expert neural classifier of the Local Linear Map – type (LLM-net) is attached [19]. The LLM-nets perform the final mapping $p_l(x) \rightarrow y \in \mathbb{R}^N$. The LLM-network is related to the self-organizing map [12] and the GRBF-approach [15]. The LLM-net performs a vector quantization using $N_{LLM}$ nodes in the input space and can be trained to approximate a nonlinear function by a set of locally valid linear mappings which are attached to the reference vectors. For details see e.g. [19].

The three processing stages are trained successively, first vector quantization and PCA-nets (unsupervised), finally the LLM-nets (supervised). Classification of input $x$ is carried out by finding the best match reference vector $r_{n(x)}$, then mapping $x$ to $p_{n(x)}(x)$ by the attached PCA-net and finally mapping $p_{n(x)}(x) \rightarrow y$. The major advantage of the VPL-classifier is its ability to form many highly specific feature detectors (the $N_{VQ} \cdot N_{PCA}$ local PCs), but needing to apply only $N_{VQ} + N_{PCA} - 1$ filter operations per classification. So a large set of filters storing implicit object knowledge has not to be paid by a high computational effort in application.

3 Application to computer vision tasks

A system that is to be used on a variety of tasks is not allowed to have a large parameter space that has to be sampled before successful application. For the
VPL-classifier, the crucial parameters are those of the local PCA \( N_{VQ}, N_{PCA} \) and the number of nodes in the expert LLM-nets \( N_{LLM} \). Search in this parameter space is easy because the system performance is well behaved in all three parameters, i.e. the classification rate (rate of correct classifications) rises continuously when the parameter values are increased until saturation is reached. This could be shown in [8] for different sample tasks.

The bandwidth of the VPL-classifier will be demonstrated on tasks of different character: (1) classification of complete objects, (2) segmentation of partially occluded objects by classification of small image windows, (3) classification of hand gestures to show the capability to deal with nonrigid objects.

3.1 Object recognition

![Examples from the Columbia Object Image Library.](image)

To evaluate the performance of the VPL-system for object classification the Columbia Object Image Library (COIL)\(^1\) was used. A description of the database is given in [17]. It consists of single objects located on a turntable in a normalized position. Each object is rotated on a turntable at pose intervals of 5 degrees, so there are 72 images of each object (RGB). Fig. 2 shows some examples.

For the test, 60 objects were used. The resolution was subsampled by a factor 2 to \( 64 \times 64 \). Since the objects are centered, the “where problem” had not to be solved here and the images could be directly used as input vectors \( x \in \mathbb{R}^{64^2} \). 18 images of each object at pose intervals of 20 degrees were used for training, the remaining 54 images for testing. The rate of correct classifications was about 93\% for a VPL with \( N_{VQ} = 9, N_{PCA} = 10 \) and \( N_{LLM} = 30 \).

3.2 Object segmentation

Partial occlusion is a difficult problem in object recognition. A key ability in the construction of algorithms that can deal with occlusion is the segmentation

---

\(^1\) Available at [http://www.cs.columbia.edu/C\&VE](http://www.cs.columbia.edu/C\&VE)
of the different object regions as it allows to detect the object borders. This is a prerequisite for subsequent classification algorithms that can mask occluded object regions.

The VPL-classifier can be used to scan an entire image using an input window that is small compared to the objects but large enough to evaluate colour texture features [21,7]. As in the previous section, the output is the object class. For the test, 20 objects of COIL were used (RGB). The system was trained with 18 images of each of the un-occluded objects, from which training windows of size $17 \times 17$ were sampled from the object area (image resolution $128 \times 128$). The dimensionality of the input vector $x$ is $3 \cdot 17^2$.

The trained system was tested on images which were artificially generated from the remaining 54 images of each object. Each test image shows two different objects that are partly overlapping, see Fig. 3. The output classes were supplemented by an additional class for the background (in total 21). A segmentation example is shown in Fig. 3 (right), indicating the resulting class of each pixel using false colours. Of all pixels, 82% could be classified correctly. This rate is naturally below the results of the previous section as the input information is much smaller.

### 3.3 Gesture recognition

Gesture recognition is usually treated as a domain apart from object classification. However, since the major benefit of ANN-based approaches is the possibility to train a system just by providing labeled views, different hand poses can be handled in the same way as different objects. Fig. 4 shows six hand gestures of a human-machine interaction scenario which were evaluated for a test. The hand is over a mouse pad and indicates six pointing directions: Left, right, forwards, backwards, up and down. For the latter two "symbolic" hand postures were defined. The camera is located above the hand. The use of the mouse pad is to
Fig. 4. Gesture classes (from upper left to lower right): “Left, right, away, towards instructor, up, down”.

indicate to the user the area where the hand must be. Though the mouse pad is cluttered background, it does not disturb recognition as the system is adapted to its structure during the training phase.

The system was tested on a series of 48 images, 8 of each gesture. Using the “leaving one out” method for evaluation, 89% correct classifications could be reached by a VPL with \( N_{VQ} = 4 \) reference vectors, \( N_{PCA} = 6 \) local PCs and \( N_{LLM} = 20 \) LLM-nodes.

4 Conclusion and Acknowledgement

Using three different scenarios, it could be shown that the VPL-classifier is well suited for view based solutions of computer vision tasks. The system is easy to adjust and has the potential for further development. The possible future development is worth some discussion.

It should not be overlooked that the view based approach still has its limitations. For example, the gesture recognition outlined in section 3.3 requires a fixed illumination. If the spectrum or direction of the illumination changes, the recognition results are much worse. Of course, different a priori known illuminations could be trained using sample images, but the system still has no inherent
invariance properties of this kind. The same argument is true for other desirable, but not yet achievable robustness or invariance features such as independence of the camera position, scale or different skin colour.

There are promising approaches to overcome these limitations, e.g. the integration of view based recognition into hybrid systems involving high level knowledge. But even with the current restrictions, view based systems can do a lot; the major problem is to find appropriate tasks. Still, most of the work in the field of visual classification is aimed to discriminate classes which correspond to isolated objects. However, there are many other visual categories which can be classified without a decomposition into single objects. An example is given in [9], where the VPL-classifier is used to judge the stability of grasping positions of a robot hand on objects. In this scenario, a decomposition would be far too difficult, thus the problem is intractable by "classical" computer vision. Nevertheless, the view based system can be trained to judging the situations in the same way as classifying isolated objects. The conclusion can be drawn that the future potential of view based, trainable recognition systems lies not only in better methods, but also in the choice of the tasks.

I would like to thank Helge Ritter for initiating the research project and raising the money as well as for many helpful discussions. This work was supported by the Deutsche Forschungsgemeinschaft (DFG) within research project SFB 360 "Situated Artificial Communicators".

References