Head-pose estimation from facial images with Subspace Neural Networks

J. Bruske, E. Ábrahám-Mumm, J. Pauli, G. Sommer
Computer Science Institute
Christian-Albrechts University zu Kiel, Germany
email:jbr@informatik.uni-kiel.de

Abstract

In this paper we describe an experimental system based on artificial neural networks (ANN) for estimating the pose of human heads from facial images. The networks utilize an efficient local subspace construction method based on optimally topology preserving maps (OTPMs) and local principal component analysis (PCA). The estimation accuracy is very high (about 1°), yet for real world applications the image preprocessing stage needs to be improved with respect to a more robust segmentation of face regions.

1 Introduction

Estimating the pose of human heads (i.e. pan and tilt) from camera images is an important task with applications in e.g. driver surveillance and the design of advanced human-machine interfaces. In this article we use subspace neural networks for solving this task. They exploit the fact that pictures from human heads of a single person that only differ in pan and tilt lie on a low- (2-) dimensional trajectory (submanifold) in image space.

Utilizing an efficient local subspace construction method based on optimally topology preserving maps (OTPMs) and local principal component analysis (PCA) we have shown in [BS97] how local subspaces can be constructed in time $O(n + m(d)^3)$, where $m(d)$ is a function of the intrinsic dimensionality d of the manifold only and n is the dimension of the embedding space. Due to this linear instead of cubic scaling with n the method becomes applicable even for very high dimensional input spaces as frequently encountered in computer vision.

In our application we use subspace variants of Ritter's Local Linear Map (LLM) [RMS91] to approximate the mapping from facial images to head-poses. We will review the subspace construction procedure in section 2, describe the Subspace-LLM in section 3 and give experimental results in section 4.

2 Efficient local subspace construction

We will now briefly review the basic procedure for efficient local subspace construction with optimally topology preserving maps as presented in [BS97, BS98]. Given a training set $T \subset \mathbb{R}^n$ and an integer $N > 0$, it proceeds in four stages (batch-variant) and supplies us with N sets of (orthonormal) eigenvectors $e_1^i, \ldots, e_{l_i}^i$, $i \in \{1, \ldots, N\}$, each set spanning a local subspace.

1. Generate a set of N centers $S = \{c_1, \ldots, c_N\} \subset \mathbb{R}^n$ as the output of a vector quantization algorithm working on the training set T.

2. Calculate the graph $G = (V, E)$ by

   (a) associating each center in S with a node in V, i.e.

   $$|V| = |S| \text{ and } c_i \in S \Leftrightarrow i \in V$$

   (b) for each $x \in T$, connecting the nodes associated with the best and second best matching centers, i.e.

   $$E = \{(i,j) \mid \exists x \in T \, \forall k \in V \backslash \{i,j\} : \max\{\| c_i - x \|, \| c_j - x \|\} \leq \| c_k - x \| \}$$

   G is called the optimally topology preserving map[1], $OTPM_T(S)$, of S w.r.t. T, cf. [BS97].

[1] The optimally topology preserving map is closely related to Martinetz' perfectly topology preserving map [MS94].

3. For each node $i \in V$ perform a principal component analysis of the set of m_i difference vectors $\{c_{1_i} - c_i, \ldots, c_{m_i} - c_i\}$, with $(c_{j_i} - c_i)$ the difference vectors between c_i and c_{j_i}, the center of its j-th direct topological neighbor in G. This yields a set of orthonormal eigenvectors $e^i_1, \ldots, e^i_{n_i}$ and corresponding eigenvalues $\mu^i_1, \ldots, \mu^i_{n_i}$, $n_i \leq m_i$.

4. For each node $i \in V$ exclude local eigenvectors e^i_j corresponding to very small eigenvalues μ^i_j, i.e. choose $0 \leq \alpha \leq 1$ and reject eigenvector e^i_j if $\frac{\mu^i_j}{\max_k \mu^i_k} < \alpha$.

Note that the central "trick" in step 3 is to use the difference vectors $(c_{j_i} - c_i)$ for PCA of each local subspace and not the data in a local region itself. First, the difference vectors have very low noise component orthogonal to the input manifold M (due to the noise reduction property of the vector quantizing stage), and second, the number of neighbors m_i of a node i in an OTPM does only depend on the intrinsic dimensionality d and is small for small d.

3 The Subspace-LLM

The Local Linear Map (LLM) as introduced by Ritter et al., [RMS91], has found widespread application for learning input - output mappings. The LLM rests on a locally linear (first order) approximation of the unknown function $f : R^n \to R^k$ and computes its output as (winner-take-all variant)

$$y(x) = A_{bmu}(x - c_{bmu}) + o_{bmu}. \qquad (1)$$

Here $o_{bmu} \in R^k$ is an output vector attached to the best matching unit (zero order approximation) and $A_{bmu} \in R^{k \times n}$ is a local estimate of the Jacobian matrix (first order term). Centers are distributed by a clustering algorithm.

Due to the first order term, the method is very sensitive to noise in the input. With a noised version $x' = x + \eta$ the output differs by $A_{bmu}\eta$, and thus the LLM largely benefits from projecting to the local subspace, cancelling the noise component of η orthogonal to M. Equally important, instead of adapting and storing $k \times n$ parameters with each matrix A_i, by first projecting to the local l_i-dimensional subspaces only matrices $A'_i \in R^{k \times l_i}$ need to be stored. This results in much better scaling with the input dimension n and, because of the reduced number of free parameters, better learning and generalization properties.

The Subspace LLM (S-LLM) proposed in this article takes the form

$$y(x) = \sum_{i \in Nh(bmu) \cup \{bmu\}} [A'_i E_i (x - c_i) + o_i]\, h_i(z_x) \quad \text{with} \qquad (2)$$
$$z_x = (x - c_i)^T E_i E_i^T (x - c_i),$$

where $E_i = [e^i_1, \ldots, e^i_{l_i}]$ denotes the local projection matrix as calculated by the efficient subspace construction procedure, $h_i(z)$ is a radial basis function and $Nh(bmu)$ the node set consisting of the direct topological neighbors of the best matching unit w.r.t. the OTPM. It not only extends the LLM by projecting onto but also by interpolating within the subspace. As basis functions we use normalized Gaussians

$$h_i(z) = \frac{\exp(-z^2/\sigma_i^2)}{\sum_{j \in Nh(bmu) \cup \{bmu\}} h_j(z)} \text{ with}$$
$$\sigma_i^2 \sim \frac{1}{|Nh(i)|} \sum_{j \in Nh(i)} (c_j - c_i)^2.$$

Batch training of the S-LLM involves optimization of the output vectors o_i and Jacobian matrices A_i by singular value decomposition (SVD) [PTVF88] as well as adaptation of the centers c_i by an incremental version of the LBG vector quantizer [LBG80].

4 Experimental results

We recorded 500 images of a doll-head (see figure 1) for the training set. Doll instead of human heads where employed because a doll head can be more easily and accurately positioned (using a robot arm). The pan of the head varied from $-75°$ to $75°$, the tilt between $-30°$ and $30°$.

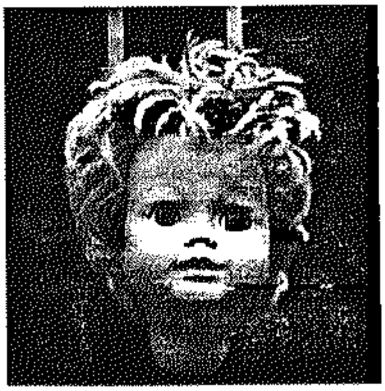

Figure 1: Left: Head of the doll hold by a robot gripper (vertical stripes at top).

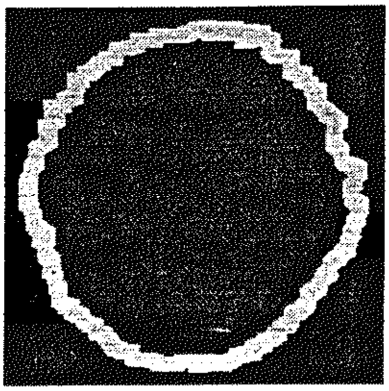

Figure 2: Reconstructed viewing circle as estimated by the S-LLM. The circle indicates the intersections of the doll's viewing directions with the computer monitor.

Preprocessing of the images involves segmentation of the face region (using color histogramming for identifying skin-colored regions), scale normalization and convolution of the grey-scale images with 64 Gabor filters. The latter are distributed on a 4 × 4 grid, with 4 orientations (0, $\pi/4$, $\pi/2$ and $3\pi/4$) on each position. As band-pass filters the Gabor filters serve two purposes: First, they filter out high frequencies which would lead to discontinuous trajectories in image space and, second, by filtering out the very low frequencies they become less sensitive to changes in brightness. The 64 filter responses serve as input to the networks, the corresponding pan and tilt angles as training output.

On a test set of 180 images (different from the training set) the Subspace-LLM achieved an averaged pose estimation error of as small as 0.64° (maximum test error 1.88°) with just 40 nodes in the network. To illustrate this result, regard figure 2. Here a doll whose pre-programmed viewing direction describes an exact circle on a computer monitor it looks at has been observed by the camera. The figure shows the circle as reconstructed from the pan and tilt angles associated with the camera images by the Subspace-LLM.

5 Conclusion

Utilizing an efficient subspace construction procedure we have presented a neural architecture, the Subspace-LLM, for estimating the head-pose from facial images with high accuracy. Results are very promising, yet for application to pose estimation of human faces in arbitrary environments we have to improve the preprocessing stage with respect to brightness invariance and more robust skin-color segmentation. The Subspace-LLM has been successfully tested in other applications as well, including an appearance based robot grasping system.

References

[BS97] J. Bruske and G. Sommer. Topology representing networks for intrinsic dimensionality estimation. In *Proc. ICANN'97*, Springer LNCS, Nr. 1327, pages 595–600, 1997.

[BS98] J. Bruske and G. Sommer. Intrinsic dimensionality estimation with optimally topology preserving maps. *IEEE PAMI*, 20(5):572–575, 1998.

[LBG80] Y. Linde, A. Buzo, and R. Gray. An algorithm for vector quantizer design. *IEEE Transaction on Communications*, 28(1):84–95, 1980.

[MS94] T. Martinetz and K. Schulten. Topology representing networks. In *Neural Networks*, volume 7, pages 505–522, 1994.

[PTVF88] W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery. *Numerical Recipes in C - The Art of Scientific Computing*. Cambridge University Press, 1988.

[RMS91] H. Ritter, T. Martinetz, and K. Schulten. *Neuronale Netze*. Addison-Wesley, 1991.