Modeling of Pose Effects in Oriented Filter Responses for Head Pose Estimation

Ilkka Kalliomäki and Jouko Lampinen

Laboratory of Computational Engineering, Helsinki University of Technology

Abstract. We propose an approach for view angle invariant recognition of 3D objects, based on modeling the variations of local feature values as function of view angle. In recognition stage we can compute the probabilities for any pixel that there is certain feature in a given pose angle. Any maximum likelihood or posterior based estimation methods can then be applied to infer the objects and their view parameters. We demonstrate the method with piecewise linear model for the pose effects, to recognize the location and pose of a head from the two eyes.

1 Introduction

View angle invariant recognition of three dimensional objects is one of the basic tasks in computer vision, and despite of vast amount of research on the subject there are no generally received solutions for the problem. Typical approaches are view invariant features, which are rather limited in recognition capacity and do not give the necessary information about the view angle, and model based fitting of the object model on the image, which lead to tedious numerical optimizations. In this paper we propose an approach, where the effect of view angle (or pose) in the feature space is modeled using simple regression methods, and in the recognition stage the view angle is inferred from one or more salient features.

The appearance of features in the human face, such as eyes and mouth, vary characteristically depending on head pose. The mouth, for example, appears mainly as a horizontal line in a directly frontal pose, but when the head is rotated, the orientation of the mouth and thus also the responses of features, such as oriented filters, change. Our goal is to build a model for the change of appearance in facial features in order to recognize facial features and pose parameters in arbitrary pose.

Subspace methods such as PCA have been commonly applied to modeling identity variation of faces in known pose with good results [1]. As the subspace of identity variation is of unknown dimensionality and nontrivial to parameterize, the PCA approach is easily justifiable. Subspace methods can be applied also to modeling pose effects in features, although a single linear subspace is insufficient, and nonlinear subspaces give significantly better results [2]. However, the dimensionality of the pose subspace is known to be exactly three, since it is spanned by three rotations, and it can be fully parameterized by for example Euler angle or quaternion representations. As an alternative to subspace modeling, one can

J. Bigun and T. Gustavsson (Eds.): SCIA 2003, LNCS 2749, pp. 156–162, 2003.

model pose variation directly in this latent space by building nonlinear regression models for oriented filter responses. Here, we consider only 2D rotations and parameterize them with azimuth and elevation angles (θ, ϕ), which act as latent variables, and the response of each filter in an oriented filter bank is modeled by a function $f_k(\theta, \phi)$.

The third rotation, parameterized for example as rotation about the view axis, is easier to model. If we use a rotationally symmetric filter bank, in which the filters in a single scale are rotated versions of each other, rotating the filters is equivalent to rotating the image about the view axis. With a sparse filter bank, the filters themselves withstand small rotations about the view axis.

2 Measuring filter responses

In this work we use Gabor filters [3] as the recognition features, with three scales and six orientations. The method proposed here can be used with any spatial filters, such as steerable filters [4] or derivative of Gaussian filters. The scales of the filters are designed in octave spacing and are chosen to be quite small so that they measure variation mainly in the orientation of local features and disregard global variation such as shading. The smallest filters have a nonzero impulse response in an approximately five pixel radius.

In order to measure the filter responses we use a synthetic head model[1]. The reference head model is deformed to match the feature locations in a frontal photograph, and texture mapped [5]. The deformation process was manually guided in order to achieve best possible visual quality.

We track feature locations in the synthetic face model for varying azimuth and elevation angles, and store the responses of filters. The filters remain centered to the feature locations as the head rotates. We have used 23 feature locations in the inner face in our experiments. The rotation angles are evenly sampled in a rectangular grid, with a total of 1500 different head poses. Fig. 1 shows a rendering of the head model with white markers added to feature locations for visualization, and zoomed left eye in several orientations.

Alternatively, it would be possible to take a large number of photographs from a real head instead of using a synthetic head. However, there are many advantages of using synthetic data to build the filter response model. The most important in practice is that measuring the filter responses from synthetic data takes far less time. Instead of taking hundreds of photographs in varied poses, the head is rendered using efficient 3D graphics hardware, and the filter responses are computed from the rendered image. Also, head pose and lighting conditions can be accurately controlled. Reliable control over pose angles is quite difficult to achieve in real-world photography. The tracking of feature locations is also easy and precise using a synthetic model. With real-world image data one must either label the feature locations manually or track them automatically.

Compared to real-world data, the main disadvantage is that the visual quality of the synthetic model is lower. The model has been built using a single frontal

[1] Shape model courtesy of University of Washington

Fig. 1. Left: Synthetic head model. Right: Tracking the center of the left eye across the pose space. The principal orientation of edges in the eye images changes considerably due to pose.

photograph, and its features become somewhat unrealistic, especially in highly rotated poses. Also, the used Phong lighting model gives rather unnatural results for human skin, and lacks cast shadows due to self-occlusions.

3 Piecewise linear models for responses

Having obtained the filter response data, we need to model it as functions $f_j(\theta, \phi)$, where θ and ϕ are the azimuth and elevation angles of the pose and i refers to filter index. This is a typical regression problem. Fig. 2 illustrates the modeling setup. In Fig. 2 *a*) a single feature, the center of the left eye, is tracked. The responses of a single oriented filter in Fig. 2 *b*) are recorded in the whole pose space in an evenly sampled grid.

Fig. 2 *c*) shows the measured amplitude responses of the filter in Fig. 2 *b*) tracking the features in Fig. 2 *a*). Large amplitude responses are obtained when the filter correlates strongly with the image. This includes a large area in the left half-plane. Fig. 2 *e*) shows the measured phase responses of the same filter. The sharp discontinuity at approximately $\theta = 40°$ occurs when the azimuthal rotation causes the center of the eye to reach the edge of the head.

We cover a part of the two-dimensional pose space with a rectangle $\theta \in [-50°, 50°]$, $\phi \in [-30°, 30°]$, with the origin of the space corresponding to a frontal view. We divide the space into 28 pieces, each piece covering approximately 15-by-15 degrees in the pose space. The piece boundaries are not optimized, but fixed, in order to simplify the modeling process.

A flexible model is needed to capture the highly nonlinear effects in the amplitude and phase regression functions. We have used piecewise linear models with fixed piece boundaries, where the complex response of an oriented filter is modeled with

$$j_k(x; a, b) = a^T x \, e^{ib^T x}, \tag{1}$$

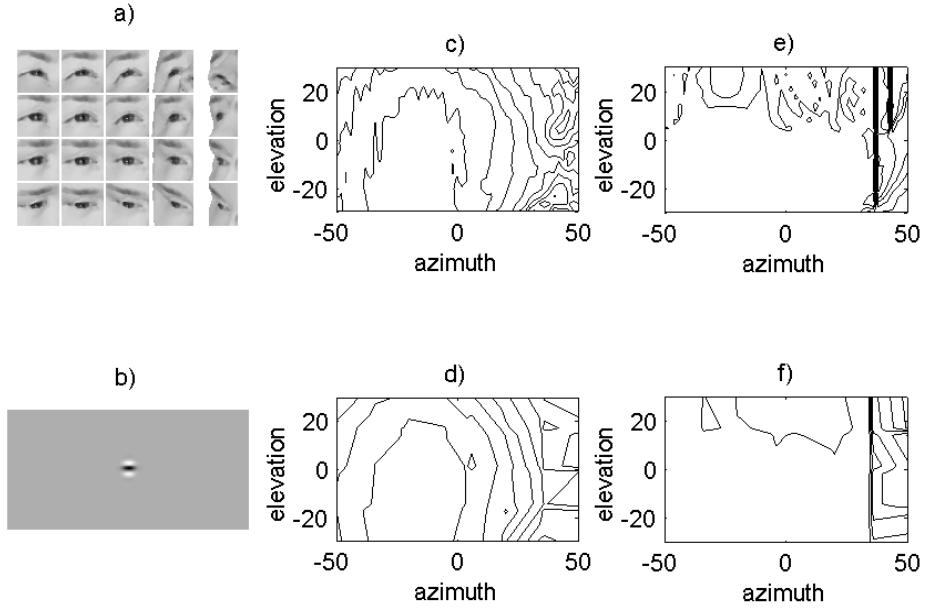

Fig. 2. a) Samples from the pose space of the feature (center of the left eye). b) A horizontally oriented filter. c) Measured amplitude response of the filter. d) Modeled amplitude. e) Measured phase response of the filter. f) Modeled phase. See text for explanation.

where $x = [\theta\, \phi\, 1]^T$ is the pose angle vector and a are the linear model parameters for amplitude and b for phase, respectively, and k is the filter index. The prediction of the whole filter bank response is obtained by stacking the models into a vector $J = [j_1, \ldots, j_n]^T$.

Amplitude responses in the measurements are typically quite smooth and the modeling process is straightforward. The model is fitted simply by computing the pseudoinverse solution which minimizes the square error between the model predictions and data. On average, the residual variance of the amplitude model is 8% of that of the original signal. Fig. 2 *d*) shows the predictions of the piecewise linear model for amplitude.

The residual variance of the phase model using original data is 30%. Unwrapping the phase by changing jumps that are larger than π to their 2π complement improves the predictions slightly. Interestingly, we found that it is beneficial to unwrap the phase along the elevation angle, the error residual dropping to 20%. Unwrapping the phase along the azimuth angle actually increased the residual to 37%. This is probably due to asymmetry in horizontal and vertical patterns in human faces. Fig. 2 *f*) shows the predictions of a piecewise linear model for phase. The discontinuity near $\theta = 40°$ occurs near a piece boundary, and the predictions of this model are fairly accurate.

4 Likelihood function

Denote a measured filter bank response vector with $J \in \mathbf{R}^{18}$ and the predicted one with J', and the normalized vectors $\mathcal{J} = J/||J||_2$ and $\mathcal{J}' = J'/||J'||_2$, respectively. We use the squared Euclidean distance between the normalized vectors,

$$\begin{aligned} S(\mathcal{J}, \mathcal{J}') &= ||\mathcal{J} - \mathcal{J}'||_2^{\,2} = (\mathcal{J} - \mathcal{J}')^T(\mathcal{J} - \mathcal{J}') \\ &= \mathcal{J}^T\mathcal{J} - 2\mathcal{J}^T\mathcal{J}' + \mathcal{J}'^T\mathcal{J}' = 2 - 2\mathcal{J}^T\mathcal{J}', \end{aligned} \tag{2}$$

as a similarity function. This is equivalent to the phase-sensitive similarity function with zero displacement in [6], written in component form as

$$S(J, J') = \frac{\sum_k |J_k||J'_k| \cos(\arg(J_k) - \arg(J'_k))}{\sqrt{\sum_k |J_k|^2 \sum_k |J'_k|^2}}. \tag{3}$$

The range of the similarity function is $[-1, 1]$, with the maximum occuring when both magnitudes and phase angles of all filter responses are equal.

The feature likelihood function is defined as the exponentiated similarity,

$$p(\mathcal{J}|\mathcal{J}') \propto e^{-\frac{1}{2}\beta||\mathcal{J}-\mathcal{J}'||_2^{\,2}} \propto e^{\beta\mathcal{J}^T\mathcal{J}'}, \tag{4}$$

where $\beta > 0$ is a constant which affects the steepness of the likelihood function and is analogous to the precision (inverse of variance) of a Gaussian distribution.

5 Results

In order to evaluate the model, we compute the likelihood functions of features for a set of new test images. The top row of Fig. 3 shows the test poses where the piecewise linear models for eye features are evaluated. The second row shows the likelihood functions of left eye feature for various view angles, and the third row for right eye feature, respectively. The likelihood functions of single features are typically multimodal, due to false matches to similar looking patterns. Obviously, the pose cannot be determined from one feature, when it is not known where the feature appears in the image. The joint likelihood of several features takes into account the spatial relations of the feature locations. To compute the joint likelihood, we shift the feature likelihoods in the (x, y)-plane corresponding to the displacement of expected feature locations in a given pose and convolve them with small Gaussian kernels in order to account for positional uncertainty in feature locations.

Assuming that the shifted features are independent, the joint likelihood is obtained by multiplying the shifted single feature likelihood functions. The joint likelihood is shown in the fourth row of Fig. 3, with origin of the two-eye template in the center of the left eye. The peak of the likelihood function is highest for the pose which is approximately correct. Also incorrect poses exhibit a lower peak in the correct location.

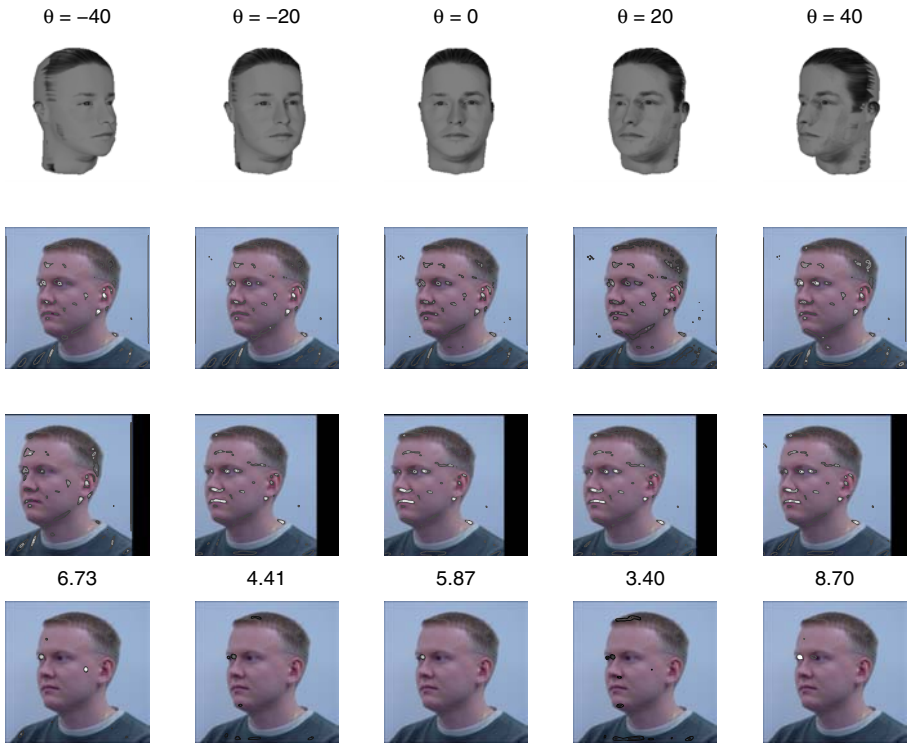

Fig. 3. Example of estimating feature location and pose. Columns corresponding to likelihoods evaluated with pose angles are shown on the first row. Second row: likelihoods of left eye. Third row: likelihoods of right eye. Fourth row: joint likelihood of eye features, with the origin of the eye template on the left eye. The maximum values of the joint likelihoods are shown on top of the images.

6 Conclusions and future work

We have presented a model for detection of facial features by explicitly modeling the variation in features due to pose as a function of rotation angles. The model facilitates also pose estimation directly using any maximization algorithm to locate the features and estimate the pose angles. Since the likelihood distributions have often multiple maxima, a global optimization method such as Metropolis sampling or simulated annealing may be prefereable.

Instead of piecewise linear models, other regression models such as polynomial, multi-layer perceptrons and kernel regression models may be used to model the filter responses. Kernel models are especially appealing, since they have an intuitive interpretation: the centers of the kernels correspond to view prototypes, scattered apart in the pose space. Also the shapes of the amplitude responses seem suitable for kernel models. The discontinuous phase responses are more difficult to model, and should be given special consideration.

A considerable limitation of the current method is that only pose variation is modeled. In addition, the features contain also intrapersonal (facial expression, lighting conditions) and interpersonal (overall head shape and texture) variation. To account for these variations, the model needs to be augmented with corresponding subspaces, possibly along lines proposed in [2].

References

1. Pentland, A. and Moghaddam, B. and Starner, T.: View-based and Modular Eigenspaces for Face Recognition. Proc. of IEEE Conf. on CVPR, Seattle, WA, 1994.
2. Okada, K. and von der Malsburg, C.: Analysis and synthesis of human faces with pose variations by a parametric piecewise linear subspace method. In Proc. of IEEE Conf. on CVPR, Kauai, 2001
3. Daugman, J.G.: Complete discrete 2-D Gabor transforms by neural networks for image analysis and compression. IEEE Trans. on ASSP **36** (1988) 1169–1179
4. Simoncelli, E. P. and Freeman, W. T.: The Steerable Pyramid: A Flexible Architecture for Multi-Scale Derivative Computation. IEEE Second Int'l Conf on Image Processing. Washington DC, October 1995.
5. Kalliomaki, I. and Lampinen, J.: Feature-based inference of human head shapes. In P. Ala-Siuru and S. Kaski, editors, STeP 2002 - Intelligence, The Art of Natural and Artificial, Proc. 10th Finnish Artificial Intelligence Conference.
6. Wiskott, L., Fellous, J.-M., Krüger, N., von der Malsburg, C.: Face Recognition by Elastic Bunch Graph Matching. In Jain, L.C., Halici, U., Hayashi, I., and Lee, S.B. (eds.), Intelligent Biometric Techniques in Fingerprint and Face Recognition. CRC Press (1999)