Person Independent Head Pose Estimation by Non-Linear Regression and Manifold Embedding

Matthias Straka[1], Martin Urschler[1], Markus Storer[1], Horst Bischof[1], Josef A. Birchbauer[2]

[1] Institute for Computer Graphics and Vision
Graz University of Technology, Austria
{straka,urschler,storer,bischof}@icg.tugraz.at

[2] Siemens Biometrics Center
Siemens IT Solutions and Services, Siemens Austria
josef-alois.birchbauer@siemens.com

Abstract

This paper describes an approach to head pose estimation in passport type images with an emphasis on high accuracy for near-frontal poses as well as person independence. Two different algorithms are proposed and compared. A Histogram of Oriented Gradients (HOG) descriptor is used for non-linear regression and a Biased Manifold Embedding (BME) approach is extended to cope with multiple pose-angles. In addition, we present an approach for the creation of an artificial training database. The effectiveness of the algorithms is shown on the artificial database as well as on a publicly available dataset where the HOG based approach performs best.

1. Introduction

The estimation of the 3D pose of a human head from still images is a difficult problem but has numerous potential applications such as head pose tracking in vehicles [8], controlling virtual avatars [7] and facial image analysis in biometric images [11]. While humans learn to quickly estimate the orientation of a head very early in their life, a computer vision system has to overcome a variety of problems which have challenged researchers and scientists for decades [9]. For example, human faces exhibit an enormous variation in facial expressions and there are huge differences in faces of people with different gender, ethnicity and age. A Head Pose Estimation System (HPES) should demonstrate robustness to such factors and also to occlusion, noise, lighting and perspective distortion.

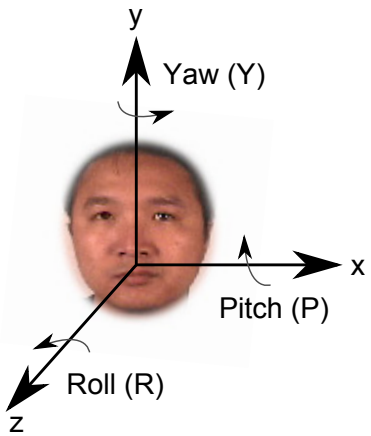


Figure 1. Definition of the human head pose angles

We are especially interested in analyzing passport type photos as defined by the International Civil Aviation Organization (ICAO) [5] in order to determine the exact head pose of the person being shown. This standard requires that the person in the image has a frontal head pose which means that the head rotation must not deviate more than ± 5 degrees in any direction from frontal. The human

head has three rotational degrees of freedom (DOF) which are shown in Figure 1. The left/right-pose is defined by the yaw angle, pitch angles describe the up/down pose and the roll angle represents in-plane rotations. Our HPES is required to estimate all of these angles accurately.

The contributions in this paper deal with enhancements of existing algorithms in order to improve their pose estimation accuracy (see Section 3.). We show how to extend the Biased Manifold Embedding [1] technique to multi-DOF pose estimation and use the histogram of oriented gradients descriptor [4] for non-linear regression [8]. In addition to that we propose a database for training a head pose estimation system. An evaluation of all methods in an unified framework allows a direct comparison of the estimation performance of different systems (see Section 4.).

2. Related Work

According to a recent survey [9] existing algorithms for head pose estimation can be categorized into several classes. *Appearance Template Methods* and *Detector Array Methods* try to estimate the head pose by either directly comparing head images with a set of template images or by using a trained detector in order to find the most similar pose for a new head image.

Regression Methods learn a continuous estimate of the head pose by a possibly nonlinear mapping of the high dimensional image features to pose angles. *Manifold Embedding Methods* assume that even though an image consists of hundreds of dimensions spanned by the pixels of the image, only a few dimensions define the pose [7]. Thus these methods map the head image to a low-dimensional manifold that is defined by the continuous pose angles. Both methods rely on a robust and accurate face localization.

Geometric Methods rely more on human perception and include measurements of distances between facial features and deviation from bilateral symmetry [13]. These methods require that facial features such as eye corners can be detected robustly. Due to occlusion, facial expressions and pose-dependent viewing angles this is a non-trivial task [12]. *Flexible Models* solve this problem by fitting non-rigid face models to the 2D image of a face in order to determine the head pose. For example, graphs of facial features can be automatically deformed until they fit to a face in an arbitrary pose [14].

3. Head Pose Estimation

The main focus of our work is the continuous estimation of the head pose in monocular still images (e.g. passport photos). This means that we require that the HPES is person independent and can handle the large variety of faces correctly. Detecting facial features like the nose or the mouth is a challenging problem and wrongly detected features will degrade the overall performance. Therefore it is beneficial if these features need not be detected in the input image.

Only two types of methods from the overview in Section 2. fulfill our requirements. Both Manifold Embedding (see Section 3.1.) as well as Non-Linear Regression methods (see Section 3.2.) perform pose angle regression and require only a good face localization, which is given in our application. In addition to that, they are able to adapt to the variety of people found in real world scenarios automatically.

We propose an HPES similar to [8] which is sketched in Figures 2 and 3. The head pose estimation starts with an image of a human head which is presented as the input to the system. If the position of

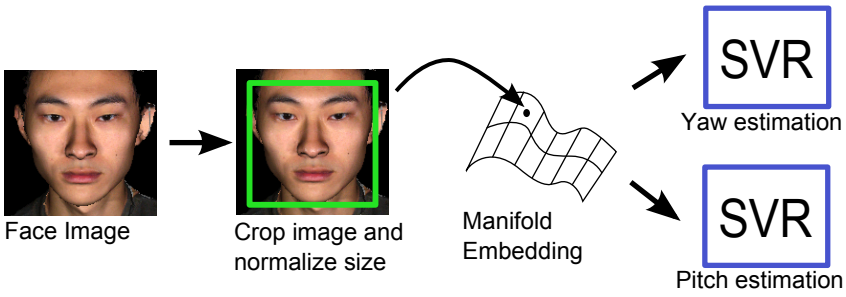


Figure 2. Components of a head pose estimation system using manifold embedding

the head in the image is not known in advance, face normalization such as [11] can be used to transform the input images such that the eyes of the subject have a fixed position in the image. Therefore the roll angle can automatically be estimated by this step. A narrow region around the face is then extracted based on the position of the eyes. The yaw and pitch angles can subsequently be estimated by learning a mapping of the pixels of the region to pose angles. The remainder of this section describes two algorithms that work within such a framework.

3.1. Biased Manifold Embedding

The first algorithm we evaluate uses a *manifold embedding* technique [1] to estimate the pose of the human head. The main idea is to map the high dimensional input image onto a low-dimensional manifold which correctly maps pose angles while ignoring lighting, occlusions and facial differences of individuals. Pitch and yaw angles can be extracted from such a manifold by using regression techniques.

We use the geometrically motivated Laplacian Eigenmaps (LE) [2] algorithm for representing high dimensional data. It allows an efficient non-linear dimensionality reduction while preserving locality properties. This makes it suitable for clustering applications such as an HPES which tries to keep similar poses within a small neighborhood on the manifold. LE require a distance measure between every tuple of data points $(\mathbf{x}_i, \mathbf{x}_j)$ from a training set such as the Euclidean distance between the i-th and j-th data point:

$$d(i,j) = \|\mathbf{x}_i - \mathbf{x}_j\| \tag{1}$$

As feature vectors $\mathbf{x}_i$ we use the facial region scaled to 32×32 pixels filtered by a Laplacian of Gaussian filter in order to extract edge information. Laplacian Eigenmaps can be generated from the distances between feature vectors alone [2]. When using the unmodified distance calculation, features of the same person at different poses are likely to have a lower distance than features of different people at the same pose, which is a disadvantage in our context. Balasubramanian et al. [1] therefore enhanced manifold embedding techniques by using a *Biased Manifold Embedding* (BME) approach that incorporates pose information into the creation of the manifold in order to weight pose differences much higher than inter-person differences:

$$\tilde{d}(i,j) = \frac{|p(i,j)|}{\max_{m,n} p(m,n) - p(i,j)} \cdot d(i,j) \qquad \text{with} \quad p(i,j) = |\phi_y(i) - \phi_y(j)| \tag{2}$$

In Eqn. (2) this biasing term is multiplied with the Euclidean distance from Eqn. (1). This ensures that

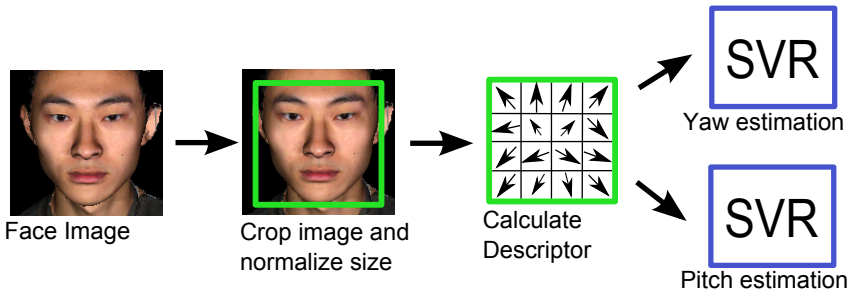


Figure 3. Components of a descriptor based head pose estimation system

data points with a small pose distance $p(i, j)$ for the images i and j lie close together on the resulting manifold. Balasubramanian et al. use only the yaw angle $\phi_y(i)$ of the head in the image. For our system we extend the definition of the pose distance by the pitch angle $\phi_p(i)$:

$$p(i,j) = \sqrt{[\phi_y(i) - \phi_y(j)]^2 + [\phi_p(i) - \phi_p(j)]^2} \tag{3}$$

The complete head pose estimation process is shown in Figure 2. With the help of Laplacian Eigenmaps, the low dimensional manifold is created from the feature vectors of several thousand training images (see Section 3.3.). We use two support vector regression machines to extract the yaw and pitch angles from the low dimensional manifold. As there exists no direct way to map new feature points onto an existing manifold, we learn the mapping from feature space to the manifold using a Gaussian Regression Neural Network (GRNN) [16] in order to estimate the pose of previously unseen head images.

3.2. Non-Linear Regression of Image Descriptors

Another approach that promises to support our requirements was initially published by Murphy-Chutorian et al. [8] who calculated a localized gradient orientation (LGO) histogram on the facial region of a normalized facial image. An LGO descriptor can be compared to a single SIFT descriptor [6] with a fixed position, orientation and scale. The system, which can be seen in Figure 3, extracts a scale normalized region around the face and calculated an LGO descriptor for it. Two support vector regression machines then estimate the yaw and pitch angle for the LGO descriptor vector. With this system pose angle estimation with a mean absolute pitch error of 5 degrees as well as 7 degrees in yaw angles are possible [8].

For our system we require a better performance and therefore propose several improvements. In order to reduce lighting effects, we enhance the shadow areas of the gray-level input image by a non-linear gamma normalization by transforming each pixel $I(x, y)$ to $(I(x, y))^\gamma$ with $\gamma = 0.2$. Instead of the LGO descriptor we use a Histogram of Oriented Gradients (HOG) descriptor [4] which has proven to be effective for object detection in images. Recent work has pointed out that HOG descriptors are well suited for head pose classification tasks [10] but they have not been used for head pose regression so far. The major advantages of the HOG descriptor over the LGO are that the orientation of the gradients are weighted according to the gradient magnitude. In addition to that the descriptor features a spatially selective normalization which stores a single histogram multiple times but with different weights. Another important improvement deals with the scale normalization of the image. While in [8] only a scaled down facial patch is used, we follow the recommendation of [4] to use the

(a) $(\phi_y, \phi_p) = (30°, -20°)$ (b) $(\phi_y, \phi_p) = (0°, 0°)$ (c) $(\phi_y, \phi_p) = (-30°, 10°)$ (d) $(\phi_y, \phi_p) = (20°, 10°)$

Figure 4. Examples of generated head poses at different yaw and pitch angles (ϕ_y, ϕ_p) and light settings

largest scale possible without smoothing. As a result, many more gradient values are gathered in the orientation histograms of the descriptor which increases its robustness.

The effects of these enhancements on the head pose estimation accuracy will be demonstrated in Section 4. Before that we present our novel face image database which allows us to train and evaluate the systems.

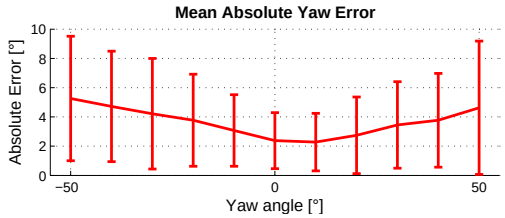
3.3. Facial Image Database

For training an image based head pose estimation system, a large database of facial images in different poses and with multiple people is needed. A variety of databases are mentioned in the literature [8, 3] but there is either not enough variation in the images or the database is not open to the public. This is why we propose a method to create a face image database with a high flexibility at a very low cost by rendering 3D models from existing laser scans. The Facial Expression database from the Binghampton University [15] contains fully textured 3D heads of 100 subjects. We generate multiple facial images from each head model by rotating it in 3D space with yaw within $\pm 50\,°$ and pitch angles within $\pm 30\,°$. The background of each rendered image is set to an arbitrary image in order to simulate a non-uniform background. In addition to that, we rotate a variable light source around the head in order to generate a variety of lighting situations. Such training data allows the system to become more robust and person independent. In all generated images the mid-point of the eyes is kept at a constant position which allows us to eliminate the face detection prior to head pose estimation. Therefore our training images are not influenced by the performance of the head localization step. Some of the synthetic images can be seen in Figure 4.

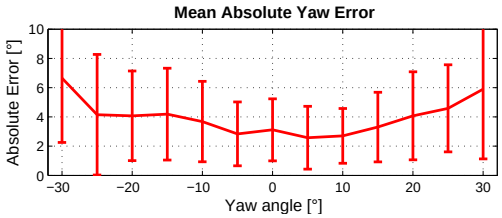
4. Experimental Results

We performed an experimental evaluation on our artificial database (see Section 3.3.) as well as on the publicly available FacePix database [3]. Three different implementations of a head pose estimation system were evaluated: Biased Manifold Embedding [1], Local Gradient Orientations mapped by an SVR [8] and our Histograms of Oriented Gradients based improvement described in this paper.

Training of the systems was performed by using a subset of head images from persons in our database while keeping the remaining persons for testing. The optimal parameters for the descriptors and machine learning algorithms were found by an exhaustive search over the parameter space. As the performance measure we used the mean absolute pose angle estimation error. This measure is often



(a) Performance on our database: $E_y = 3.6°$



(b) Performance on the FacePix database: $E_y = 4.0°$

Figure 5. Yaw angle estimation performance of the HOG descriptor with SVR learning for different yaw angles (mean and standard deviation)

used in the literature and therefore allows comparison with other publications (e.g. [9]):

$$E_y = \frac{1}{N}\sum_{i=1}^{N} |\phi_y(i) - P_y(\mathbf{D}_i)| \qquad E_p = \frac{1}{N}\sum_{i=1}^{N} |\phi_p(i) - P_p(\mathbf{D}_i)| \tag{4}$$

It is defined as the absolute difference between the ground-truth pose angle of the i-th image (e.g. $\phi_y(i)$) and the predicted angle from the descriptor $\mathbf{D}_i$ of the image. The mean over all N images in the test-set yields the final value for either yaw or pitch estimation errors.

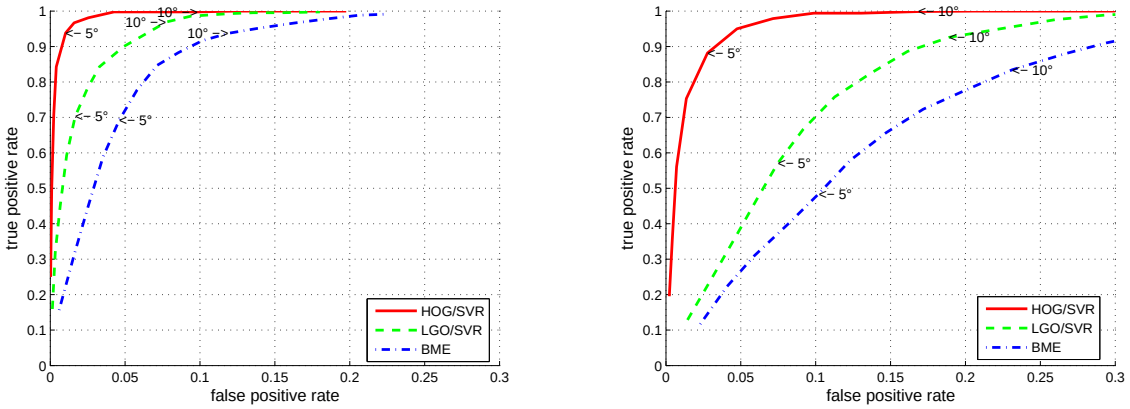
Method	$\mathbf{E_y}$ our DB	$\mathbf{E_p}$ our DB	$\mathbf{E_y}$ FacePix
Biased Manifold Embedding	5.6 °	6.5 °	14.0 °
LGO / SVR	4.5 °	4.4 °	7.5 °
HOG / SVR	**3.6** °	**3.1** °	**4.0** °

Table 1. Comparison of the Biased Manifold Embedding, LGO and HOG approach in terms of mean absolute pose angle estimation error when trained and tested on our database and additionally tested on the FacePix database [3]

In Table 1 the performance of the systems on yaw and pitch angle estimation is compared using the mean absolute error. It can clearly be seen that the HOG approach outperforms the other methods in both pitch and yaw angle estimation performance. The biased manifold embedding approach shows the worst results in our evaluation. This finding is somewhat contrary to what is stated in [1] where mean absolute errors of around 2 ° are promised. The reason for this is that the original algorithm is trained and tested on the FacePix database which only contains rotations around a single axis (left/right). When used in a multi-dimensional pose estimation system or heterogeneous datasets, this performance degrades due to ambiguities between neighboring poses.

A more detailed evaluation of the HOG descriptor based system in Figure 5(a) shows that the yaw angle estimation performance has a U-shaped curve for different yaw angles with a minimum at the 0 ° yaw angle. This means that yaw angles are estimated more accurately in frontal pose head images. Such behavior is beneficial in applications such as ours where we want to estimate the head pose in near-frontal images. A similar evaluation in Figure 5(b) shows a good performance of the HOG based system with an average yaw estimation error of 4.0 ° on the FacePix database [3] while still being trained on our artificial database.

As the ICAO standard [5] allows only frontal head poses for passport images, we evaluate the performance of the classification of head images into frontal and non-frontal poses as well. We do this by thresholding the estimated yaw and pitch angles for a given head pose image. In Figure 6 we show the Receiver Operating Characteristics (ROC) as true and false positive rates for different threshold



(a) Thresholding of yaw angles

(b) Thresholding of pitch angles

Figure 6. Evaluation of the frontal-pose classification performance of the three system using ROC plots

values on non-training images from our generated database. At a threshold of slightly more than five degree we are able to classify over 90% of all faces correctly with less than 5% false positives when using the HOG based HPES, which is due to the low estimation error in frontal poses. Other methods do not show such a good performance for frontal pose classification. We obtain similar values for the same evaluation of the FacePix database but with twice as many false positives (not shown in this paper).

5. Conclusion

In this paper we presented a system for head pose estimation from monocular still images. We estimated continuous pose angles from either localized gradient orientation (LGO) histogram descriptors, histogram of oriented gradients (HOG) descriptors or manifold embedded pixel values. The main contributions were the extension of the biased manifold embedding approach [1] towards pose estimation in two DOF and use of the HOG descriptor [4] for head pose estimation. While the HOG descriptor has become quite common in object detection, we showed that its advantages hold in regression tasks as well. The biased manifold embedding approach did not show convincing results in our experiments when used for the estimation of more than one pose angle.

For training the HPES a novel database was generated from 3D head models, also background clutter and lighting variations were simulated. We showed that a system trained on our database also performed well on different databases. In future work we intend to perform more experiments to test the method on more complex data (e.g. face with eye glasses and/or different facial expressions).

Acknowledgement

This work has been funded by the Biometrics Center of Siemens IT Solutions and Services, Siemens Austria.

References

[1] V. N. Balasubramanian, S. Krishna, and S. Panchanathan. Person-independent head pose estimation using biased manifold embedding. *Advances in Signal Processing*, 2008:1–15, 2008.

[2] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computing*, 15(6):1373–1396, 2003.

[3] J. Black, M. Gargesha, K. Kahol, P. Kuchi, and S. Panchanathan. A framework for performance evaluation of face recognition algorithms. In *Proceedings of Internet Multimedia Management Systems III*, 2002. `http://www.facepix.org/`.

[4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proceedings of the International Conference on Computer Vision & Pattern Recognition*, volume 2, pages 886–893, June 2005.

[5] ISO/IEC 19794-5:2005. Information technology – biometric data interchange formats – part 5: Face image data, June 2005.

[6] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60:91–110, 2004.

[7] S. J. McKenna and S. Gong. Real-time face pose estimation. *Real-Time Imaging*, 4(5):333–347, 1998.

[8] E. Murphy-Chutorian, A. Doshi, and M. Trivedi. Head pose estimation for driver assistance systems: A robust algorithm and experimental evaluation. In *Proceedings of the IEEE Conference on Intelligent Transportation Systems*, pages 709–714, October 2007.

[9] E. Murphy-Chutorian and M. Trivedi. Head pose estimation in computer vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(4):607–626, 2009.

[10] E. Ricci and J.-M. Odobez. Learning large margin likelihoods for realtime head pose tracking. In *Proceedings of the International Conference on Image Processing*, 2009.

[11] M. Storer, M. Urschler, H. Bischof, and J. A. Birchbauer. Face image normalization and expression/pose validation for the analysis of machine readable travel documents. In *Proceedings of the 32nd Workshop of the Austrian Association for Pattern Recognition*, pages 29–39, May 2008.

[12] T. Vatahska, M. Bennewitz, and S. Behnke. Feature-based head pose estimation from images. In *Proceedings of the IEEE-RAS 7th International Conference on Humanoid Robots*, 2007.

[13] H. R. Wilson, F. Wilkinson, L.-M. Lin, and M. Castillo. Perception of head orientation. *Vision Research*, 40(5):459–472, 2000.

[14] L. Wiskott, J.-M. Fellous, N. Krüger, and C. von der Malsburg. Face recognition by elastic bunch graph matching. In L. Jain, editor, *Intelligent Biometric Techniques in Fingerprint and Face Recognition*, chapter 11, pages 355–396. CRC Press, 1999.

[15] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato. A 3D facial expression database for facial behavior research. In *Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition*, pages 211–216, 2006.

[16] Q. Zhao, D. Zhang, and H. Lu. Supervised LLE in ICA space for facial expression recognition. In *Proceedings of the International Conference on Neural Networks and Brain*, volume 3, pages 1970–1975, Oct. 2005.