Facial Pose from 3D Data

Ajit Rajwade$^1$ and Martin D. Levine$^2$

Center for Intelligent Machines, McGill University, Montreal H3A2A7, Canada

Abstract

The distribution of the apparent 3D shape of human faces across the view-sphere is complex, owing to factors such as variations in identity, facial expression, minor occlusions and noise. In this paper, we use the technique of Support Vector Regression to learn a model relating facial shape (obtained from 3D scanners) to 3D pose in an identity-invariant manner. The proposed method yields an estimation accuracy of 97% to 99% within an error of +/- 9 degrees on a large set of data obtained from two different sources. The method could be used for pose estimation in a view-invariant face recognition system.

Keywords: 3D Pose Estimation; Range Data; Discrete Wavelet Transform; Support Vector Regression; Dimensionality Reduction; Principal Components Analysis.

1. Introduction

Although 3D face recognition systems have not received very wide attention in the literature, new technologies are likely to make them more interesting and practical in the near future. Such systems have the potential of being more accurate than current 2D systems since 3D image capture eliminates the projection uncertainties associated with head pose in conventional 2D capture. Nevertheless, head pose would still remain an issue.

The apparent 3D shape of a human face undergoes considerable change across different poses over the view-sphere and will lead to inaccurate matching in face recognition systems. It is known that an estimate of facial pose can improve face recognition accuracy considerably. In this paper, we outline a learning method to predict facial pose from a 3D scan using only shape information. The method is efficient and generic and uses the well-known technique of Support Vector Regression. We also suggest a method using the discrete wavelet transform that enhances pose-specific details to improve the accuracy of pose estimates.

The paper outline is as follows: Section 2 gives an overview of existing methods for prediction of facial pose from 2D and 3D images. Section 3 provides some background material on the method of Support Vector Regression, as well as the outline of the method employed, including the training and testing procedure. Section 4 describes the

---

$^1$ School of Computer Science
$^2$ Department of Electrical, Computer and Software Engineering
experimental results. Conclusions and directions for future research are given in Section 5.

2. Background

Several approaches have been proposed in the literature for face pose estimation from 2D images. These can be broadly categorized into feature- and appearance-based methods. The former [1] relate facial pose to spatial arrangements of significant facial features such as the eyes or nose. A common method is to calculate the facial pose from the angle subtended by the line joining the two eyes with respect to the chosen axes (for example [1]). Such methods depend heavily on the accuracy of the feature detection algorithms, which are difficult to implement in a robust enough manner. This is particularly true for a large range of poses across the view-sphere, since the apparent shape of the individual facial features themselves undergo a change with pose.

Appearance-based methods, on the other hand, consider the face as a whole entity. Pentland et al [2] and Srinivasan and Boyer [3] use methods based on PCA for estimating the facial pose. Motwani and Ji [4] precede the PCA-based pose-estimation calculation by a pre-processing stage involving computation of the wavelet sub-bands of the facial images. Similarly, Wei et al use Gabor filters for pre-processing [5]. Li et al [6] and Huang et al [7] have used Support Vector Regression and classification for the purpose of pose estimation. Li and colleagues [8] have proposed a method in which facial images are projected onto a lower-dimensional space using kernel PCA, followed by support vector classification of the lower-dimensional patterns.

In the case of 2D images, pose-estimation methods need to be robust to changes in illumination. 3D range data are independent of illumination, which makes the pose-estimation problem ostensibly simpler. Nevertheless, the distribution of facial range images across the view-sphere is still quite complex owing to differences in identity, facial expression, noise or minor occlusions. Prior research on 3D-pose estimation from range data includes a PCA-based method [9], which is essentially a tracking algorithm. It estimates pose from a sequence of scans of a person rotating his head, making use of information from previous scans in the sequence. Another approach is discussed in [10], where an ellipsoid is first fitted to a cloud of points in 3D. The head-pose is then calculated from the orientation of the axes of the ellipsoid with respect to the chosen axes.

In this paper, we assume that there is an inherent similarity between the 3D shape of faces of different individuals in similar poses and use a learning method to exploit this. We use a combination of the Discrete Wavelet Transform [11] and Support Vector Regression [12] to arrive at a generic relationship between faces and pose, the latter being defined in terms of the angles of rotation around the Y- and X-axes. To our knowledge, ours is the first attempt to develop such a generic relationship between pose and 3D facial shapes in different poses. Our method is specifically designed to predict the pose from a single 3D scan of a person, making use of only the model learnt by the Support Vector Machine.
3. 3D Pose Estimation

3.1 Theory: Support Vector Regression

Support Vector Machines have emerged as a powerful classification and regression technique in computer vision [13, 14]. They are based on the principle of structural risk minimization [12]. Consider a set of \( l \) input patterns denoted as \( x \), with their corresponding class-labels, denoted by the vector \( y \). A support vector machine obtains a functional approximation given as 
\[
f(x; \alpha) = w \cdot \Phi(x) + b,
\]
where \( \Phi \) is a mapping function from the original space of samples onto a higher dimensional space, \( b \) is a threshold, and \( \alpha \) represents a set of parameters of the SVM. If \( y \) is restricted to the values \(-1\) and \(+1\), the approximation is called support vector classification (SVC). If \( y \) can assume any valid real value, it is called support vector regression (SVR). By using a kernel function given as 
\[
K(x, y) = \Phi(x) \cdot \Phi(y),
\]
the problem of support vector regression can be modeled as the following optimization problem:

Maximize 
\[
W(\alpha^*, \alpha) = -0.5 \sum_{i,j=1}^{l} (\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)K(x_i, x_j) - \varepsilon \sum_{i=1}^{l} (\alpha_i^* + \alpha_i) + \sum_{i=1}^{l} y_i(\alpha_i^* - \alpha_i)
\]

Subject to the following conditions:
\[
\sum_{i=1}^{l} (\alpha_i^* - \alpha_i) = 0 \quad \text{and} \\
0 \leq \alpha_i^*, \alpha_i \leq C
\]

Here \( \varepsilon \) denotes the regression error. The solution to this problem is given by the equation 
\[
f(x) = \sum_{i=1}^{l} (\alpha_i^* - \alpha_i)K(x, x_i) + b.
\]
In most cases, only a small fraction of the training samples have non-zero values of \( \alpha \). It is solely these examples that influence the final decision function, and these are referred to as the support vectors. The significance of using kernel functions is that they perform an implicit (and efficient) mapping onto a higher dimensional space and improve the separability between the different classes of the training data [12].

3.2 Collection and Processing of Data

The proposed pose-estimation process consists of off-line training followed by a testing phase. The data for the experiments was collected from two sources, the first being the set of eigenvectors of manually aligned 3D shapes of human faces, provided by the University of Freiburg [15]. As per the morphing method explained in [15], 100 new faces were obtained. The morphing method consists primarily of taking a linear combination of the provided eigenvectors to generate new shapes. The coefficients for the linear combination were chosen randomly within the prescribed range of +/- 3 times
the corresponding eigenvalues [15]. The faces thus morphed were in exact frontal pose (0
degree head rotation about either axis). Since the faces obtained in this way were in the
form of point-clouds, a surface reconstruction step was necessary. This was performed
using the ‘Power Crust Surface Reconstruction Algorithm’ [16]. To obtain training data
in all possible poses, the facial surfaces were suitably projected onto different view-
planes across the view-sphere using the well-known Z buffer algorithm [17]. This is
actually equivalent to rotating the facial surface, but is much more efficient to implement.
The view-sphere was suitably sampled so as to obtain all views of the face corresponding
to combined rotations from 0 to +90 degrees around the Y- and from -30 to +30 degrees
around the X-axis in steps of 3 degrees.3

The second source of data was the facial range image database from Notre Dame
University [18], which contains near-frontal range images of 277 individuals. An initial
step that involved manual alignment of the range images was required to create exact
frontal poses. For this the positions of the eyes were marked manually, and the line
joining the eyes was aligned with the horizontal. Range images that contained holes
(missing data) were passed through a simple averaging filter. Portions of the images
exterior to the facial contour were manually cropped. Different poses of each face were
generated by projection onto different view-planes, as described above.

Finally, all range images were resized to 160 by 160, taking care to preserve the aspect
ratio and padding an appropriate number of zeroes. Figures (1) and (2) show a few faces
from the Freiburg and Notre Dame databases, respectively.

3.3 Use of the Discrete Wavelet Transform

A discrete wavelet transform of level three was performed on all the training images.
Each level of wavelet decomposition yields four sub-bands, each of size equal to one-
fourth of that of the original range-image. The four sub-bands are LL, LH, HL and HH,
corresponding to the low or high frequency components across the rows or columns. The
main advantage of the discrete wavelet transform is that it provides a principled way of
downsizing the range images. Of the four sub-bands, only the LL sub-band of the range
image was used. As the LL sub-band contains only the low-frequency components of the
image, it is relatively free of noise in comparison to the other three sub-bands.
Furthermore, low frequency information is known to accentuate pose-specific details,
suppress individual facial details, and be relatively invariant to facial expressions [4]. As
will be seen, our results confirm this notion. Convolving the images with Gabor wavelets
is also another method of accentuating pose-specific information, as reported in [5].
However this method is computationally expensive and also requires careful selection of
various Gabor wavelet parameters such as center frequency, scale, kernel-size and so on.

The Haar wavelet was used in all experiments described in this paper. Tests with other
types of wavelets did not yield any significantly different results.

---

3 As the human face is symmetrical about the Y-axis, the poses from 0 to -90 degrees around the Y-axis
were not considered.
3.4 Training using Support Vector Regression

For the purpose of training, two different SVMs were used in the pose estimation experiments. One was for learning a relationship between range images and their Y-angle, and the other for learning the relationship between range images and their X-angle. The available data were divided into non-intersecting training and test sets. The training data consisted of all poses from 0 to +90 degrees around the Y-axis and 0 to +30 degrees around the X-axis in steps of $\delta$ degrees. Fifty individuals each from the Freiburg and Notre Dame databases were selected. The rest of the faces were used for testing. Two different SVM-based estimators were developed using the LIBSVM package [19]. A radial basis function kernel given by $K(x, y) = \exp(-\gamma|x-y|^2)$ with parameters $\gamma = 0.03125$ was chosen for both estimators. The parameter $C$ for the SVM was selected to be 64, and the value of the regression-margin $\varepsilon$ was chosen to be 1.0. These values were found automatically by means of cross-validation on the training data using a simple “grid-search” tool provided within the LIBSVM package [19]. The original range images of size 160 by 160 were converted to sub-bands of size 20 x 20 after level-3 wavelet decomposition. Each range image was thus represented as a 1 x 400 vector. These vectors, labeled by their pose, were given as input to the SVMs. The number of support vectors was observed to be approximately 12% and 14% of the number of training samples when creating a functional approximation for the Y- and X-angle, respectively.

3.5 Testing

The functions yielded by the SVM were tested on all of the different poses of the test faces. The test images were also decomposed using a level 3 discrete Haar wavelet transform and the LL sub-band was given as input to the SVM for testing. The test set always consisted of individuals different from those in the training set. To confirm the stability of the approach, the individuals in the training and test sets were randomly exchanged. The experiments were repeated over 30 times. The pose-estimates were compared with the known ground-truth values of both the Y- and X-angles in every single run.

4. Experimental Results

4.1 Angle estimates

Initial experiments were performed with the angular sampling size $\delta$ to determine its optimum value. Figure (3) illustrates the variation of regression accuracy with respect to $\delta$. Clearly, the mean error in pose estimation bears a direct linear relationship to the angular sampling. A value of $\delta = 3$ degrees provides the best performance. Thus, in all the experiments reported below, angular sampling is set to 3 degrees to obtain as accurate a pose model as possible, albeit at the cost of greater training time and required storage. Smaller values of $\delta$ did not improve the performance much further (as seen in Figure (3)) and the training time and storage were significantly higher.
Table 1 summarizes the average pose-estimation results obtained over all 30 runs. It should be noted that the results did not vary widely across the 30 different runs. This confirms the stability of the model.

The histograms of estimation error versus head pose angle are shown in Figures (4) and (5). For the Freiburg database, the mean error (i.e., mean value of the absolute difference between the actual and predicted pose) is 2.8 degrees and 2.58 degrees for the Y- and X- angles, respectively. For the Notre Dame database, the mean error is 3.2 degrees and 2.72 degrees for the Y- and X-angles respectively. The mean error reported in [9] is less than 2 degrees. However, as noted in Section 2, the approach in [9] makes use of information from previous frames of a rotating head sequence. Our approach predicts the pose from a single 3D scan.

4.2 Effect of range image size

In order to examine the effect of range image scale on facial pose estimation from range data, the same experiments were performed on range images of different sizes: 320 x 320, 240 x 240, 160 x 160, 100 x 100, 80 x 80 and 64 x 64. The range images of each size were decomposed by a level-3 Haar wavelet transform, to yield patterns of size 40 x 40, 30 x 30, 20 x 20, 13 x 13, 10 x 10 and 8 x 8 respectively. Figures (6) and (7) graph the relationship between estimation accuracy and sub-band size. It is observed that sub-bands of size 20 x 20 (160 x 160 range images) yield the greatest pose estimation accuracy within +/- 9 degrees. For sizes larger than 20 x 20, individual-specific details interfere with pose estimation. Sizes smaller than 20 x 20 become progressively less adequate to clearly discriminate between poses differing by around 9 degrees.

4.3 Results on different sub-bands

The pose-estimation experiments were repeated for 20 x 20 LH, HL and HH sub-bands. The estimation accuracy with these sub-bands was always clearly less than that with the LL sub-band.

4.4 Results with Dimensionality Reduction

In the estimation phase, the time complexity of support vector regression is $O(DN_{SV})$ where $D$ is the size of the input pattern and $N_{SV}$ is the number of support vectors [12]. The speed of pose-estimation could be improved considerably if the input patterns could be projected onto a lower-dimensional space before performing SVR. To achieve this, we employed the technique of PCA on the entire set of range images in different poses. SVR was then performed on the set of eigen-coefficients. As can be observed in figures (8) and (9), the accuracy of estimation was best for a dimensionality of 40 or more, though the performance was always good for a dimensionality greater than 15. The accuracy was always slightly less than that with SVR on the LL sub-bands. As

---

4 For $\delta=3$ degrees.
5 Other methods of downsizing the images were not tested.
the dimensionality was decreased beyond 15, the number of support vectors selected
during training increased rapidly, while the resulting test accuracy decreased.

5. Conclusion

We have proposed a method for generic estimation of 3D facial pose in terms of angle of
rotation about the Y-axis and X-axis. The results obtained on a large dataset obtained
from two different sources are consistent. The main advantage of the method is that the
pose of any 3D facial scan can be predicted to a good degree of accuracy using just a
small number of training faces. Such a module could be used in a pose-invariant 3D face
recognition system to predict the pose of a test-scan with a low margin of error.
Subsequently, the faces registered in the database in a frontal pose could be rotated to the
estimated angles, and accurate face matching could then be performed. Another
interesting direction for future research could be the development of an effective
unsupervised learning method to cluster faces of different individuals based on their pose.
It would also be worthwhile to explore alternative dimensionality reduction techniques to
be able to discriminate between different poses in very low dimensional spaces.

6. Acknowledgement

Martin Levine would like to thank the Natural Sciences and Engineering Research
Council of Canada for its financial support. The authors would also like to thank the
University of Freiburg [15] and Notre Dame University [18] for providing us with 3D
face databases.

References

Symmetry Plane Using Range and Intensity Images”, 14th International
2. A. Pentland, B. Moghaddam, T. Starner, “View-Based and Modular Eigenspaces
for Face Recognition”, Proceedings of IEEE Conference on Computer Vision and
Pattern Recognition, pp. 84-91, 1994.
3. S. Srinivasan and K. Boyer, “Head-pose estimation using View-based
Eigenspaces”, Proceedings of the 16th International Conference on Pattern
Proceedings of the International Conference on Image Processing, Vol. 1, 1050-
5. Y. Wei, L. Fradet, T. Tan, “Head Pose Estimation Using Gabor-Eigenspace
Modeling”, Proceedings of the International Conference on Image Processing,
6. Y. Li, S. Gong, H. Liddell, “Support Vector Regression and Classification Based
Multi-View Face Detection and Recognition”, Proceedings of the IEEE
International Conference on Automatic Face and Gesture Recognition, pp. 300-
17. E. Catmull, “A Sub-Division Algorithm for Computer Display of Curved Surfaces”, PhD Thesis, Department of Computer Science, University of Utah, Salt Lake City, USA.
About the Author: AJIT RAJWADE was born in 1979 in Mumbai, India. He completed his Bachelors Degree in Computer Engineering from the Government College of Engineering, Pune, India in 2001 and is currently pursuing his Masters in Computer Science at McGill University, Canada.

About the Author: MARTIN D. LEVINE received the B. Eng. and M. Eng. degrees in Electrical and Computer Engineering from McGill University, Montreal, in 1960 and 1963, respectively, and the Ph.D. degree in Electrical Engineering from the Imperial College of Science and Technology, University of London, London, England, in 1965. He is currently a Professor in the Department of Electrical and Computer Engineering, McGill University and served as the founding Director of the McGill Center for Intelligent Machines (CIM) from 1986 to 1998. During 1972-1973 he was a member of the Technical Staff at the Image Processing Laboratory of the Jet Propulsion Laboratory, Pasadena, CA. During the 1979-1980 academic year, he was a Visiting Professor in the Department of Computer Science, Hebrew University, Jerusalem, Israel.

His research interests include computer vision, image processing and artificial intelligence, and he has numerous publications to his credit on these topics. As well, he has consulted for various government agencies and industrial organizations in these areas. Dr. Levine was a founding partner of AutoVu Technologies Inc. (http://www.autovu.com) and VisionSphere Technologies Inc.

Dr. Levine has authored the book entitled “Vision in Man and Machine” and has co-authored “Computer Assisted Analyses of Cell Locomotion and Chemotaxis”. Dr. Levine is on the Editorial Board of the journal “Computer Vision and Understanding”, having also served on the Editorial Boards of the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE and “Pattern Recognition”. He was the Editor of the Plenum Book Series on “Advances in Computer Vision and Machine Intelligence”. He was the General Chairman of the Seventh International Conference on Pattern Recognition held in Montreal during the summer of 1984 and served as President of the International Association of Pattern Recognition during 1988-1990. He was also the founding President of the Canadian Image Processing and Pattern Recognition Society.

Dr. Levine was elected as a Fellow of the Canadian Institute for Advanced Research in 1984. During the period 1990-96 he served as a CIAR/PRECARN Associate. He is a Fellow of the IEEE and the International Association for Pattern Recognition. Dr. Levine was presented with the 1997 Canadian Image Processing and Pattern Recognition Society Service Award for his outstanding contributions to research and education in Computer Vision.
FIGURES

Figure 1: Sample Faces from the Freiburg Database

Figure 2: Sample Faces from the Notre Dame Database
Figure 3: Pose Estimation Accuracy versus Angular Sampling (Freiburg Database)

Figure 4: Error Histogram (Y-angle, Freiburg Database)
Figure 5: Error Histogram (X-angle, Freiburg Database)

Figure 6: Effect of Input Size on Accuracy of Estimation of Y-angle
Figure 7: Effect of Input Size on Accuracy of Estimation of X-angle

Figure 8: Pose Estimation Accuracy vs. Number of Principal Components
Figure 9: Pose Estimation Accuracy (X-angle) vs. Number of Principal Components

| TABLES |
|-----------------|-----------------|-----------------|-----------------|-----------------|
|                | Results for Y-angle (Freiburg Database) | Results for X-angle (Freiburg database) | Results for Y-angle (Notre Dame Database) | Results for X-angle (Notre Dame database) |
| Number of support vectors (% of number of training samples) | 12% | 14% | 12% | 14% |
| Frequency of error less than +/-3 degrees | 70.09% | 73.23% | 66% | 69.23% |
| Frequency of error less than +/-6 degrees | 94.92% | 95.97% | 91.92% | 92% |
| Frequency of error less than +/-9 degrees | 98.85% | 99.23% | 96.86% | 98.61% |
| Average Pose Estimation error (Absolute Value) | 2.8 degrees | 2.58 degrees | 3.2 degrees | 2.72 degrees |