3D HEAD POSE NORMALIZATION WITH FACE GEOMETRY ANALYSIS, GENETIC ALGORITHMS AND PCA

VITOANTONIO BEVILACQUA*†‡, FRANCESCO ANDRIANI*§ and GIUSEPPE MASTRONARDI*†

*Dipartimento di Elettrotecnica ed Elettronica, Politecnico di Bari, Via Orabona, 4, Bari 70125, Italy
†e.B.I.S. s.r.l. (electronic Business In Security), Spin-Off of Polytechnic of Bari, Via Pavoncella, 139, Bari 70125, Italy
‡bevilacqua@poliba.it
§francesco.andriani@mail.polimi.it

In this paper, a software toolchain is presented for the fully automatic alignment of a 3D human face model. Beginning from a point cloud of a human head (previously segmented from its background), pose normalization is obtained using an innovative and purely geometrical approach. In order to solve the six degrees of freedom raised by this problem, we first exploit the human face’s natural mirror symmetry; secondly, we analyze the frontal profile shape; and finally, we align the model’s bounding box according to the position of the tip of the nose. The whole procedure is considered as a two-fold, multivariable optimization problem which is addressed by the use of multi-level, genetic algorithms and a greedy search stage, with the latter being compared against standard PCA. Experiments were conducted utilizing a GavadbDB database and took into account proper preprocessing stages for noise filtering and head model reconstruction. Outcome results reveal strong validity in this approach, however, at the price of high computational complexity.

Keywords: Head; 3D; model; face recognition; pose estimation; symmetry; genetic algorithms; PCA.

1. Introduction

Recent progress in 3D scanning technology has spurred great interest from industrial and academic worlds in 3D-related research topics such as: enhanced web semantics, online model search and retrieval (also known as shape matching), machine-man interaction and 3D face recognition, just to name a few. From a technical point of view, a tri-dimensional representation of a given object can be obtained in various ways which mainly derive from the common techniques of laser scanning and
structured light pattern projection. Data acquisition can usually be tuned with progressive levels of detail, thus setting the desired trade-off between time consumption and accuracy on a case-by-case basis.

Specifically, and for what concerns face recognition, the advent of tri-dimensional data is widely considered an effective way of overcoming the most prominent limitations of two-dimensional data driven methods, such as sensitivity to illumination changes, pose variations and background clutter. Bowyer et al. presented a comparative survey of 3DFR algorithms and concluded that 3DFR has the potential to overcome the limitations of its 2D counterpart. Also Romdhani et al. reviewed recent advances in FR algorithms, centering analysis on what has been developed to address the joint challenges of pose and lighting. The latest developments of FR look to merge both 2D and 3D content potential at feature extraction level and score/decision level. Due to their interesting features of simplicity and semantic completeness, notable effort has been sustained in the analysis of range images (also known as 2.5D depth maps) since the publication of Gordon's paper. Nevertheless, resilience of 3DFR techniques to variations in face expression and head pose, is still an open problem.

To mitigate the effect of human expressions, often considered as an intrinsic form of noise in observations, several approaches have been proposed to date. Methods that deform the original face model to mimic different expressions were explored, as what was done by Lu et al. Various weighting schemes tried to reduce the influence of expression-prone areas of the face. For instance, Spatially Weighted Hausdorff Distance and Spatially Eigen-Weighted Hausdorff Distance assigned different weighting factors for different facial regions in computing the distance for face recognition. In Lin et al. the weighting function, which reflects the discriminative properties of face edge images effectively, is based on the eigenface of face edge images. In Hu and Wang, a similarity measure based on Hausdorff distance (SMBHD) can provide not only the dissimilarity information but also the similarity information of two objects to compare them. Kanade and Yamada conducted a systematic analysis on how the discriminative power of different parts on human faces changes according to different poses, and such analysis leads to a probabilistic approach to face recognition. In Chang et al. the discriminative properties of variously selected regions around the nose, were investigated.

Seminal research was conducted on different kinds of descriptive metrics, looking for features which could, to some extent, be considered invariant to pose and/or expression modifications. A powerful shape descriptor called Point Signatures was proposed by Chua et al., which is claimed to be invariant to facial expression changes. Gabor wavelets' properties, in particular, were extensively exploited for FR in the last few years, when applied to either intensity or range images. Liu and Wechsler analyzed the behavior of ICA of Gabor features; also Vinay Kumar et al., Huang et al., Zhong et al. studied how features based on Gabor filter responses could handle the uncontrolled problems encountered in 2DFR and 3DFR (illumination changes, distorted mesh, misalignment, hair occlusion, glasses).
Finally, another wavelet-based method has been proposed by Abbas et al.\(^\text{19}\) Lin et al.\(^\text{20}\) developed a family of 2D and 3D invariant features, called summation invariant, that are invariant to Euclidean transformation in both 2D and 3D. Approaches for addressing pose variations can be largely classified into two categories. The first type of these approaches is often called multi-view face recognition, and is a sort of naive extension of frontal face case: dealing with distortion through pose change, is reduced to including images taken from different points of view within the training set/face database. Significantly interesting results have been reported in Burke and Achermann’s,\(^\text{21}\) where a nearly 100\% recognition rate on a database of 24 individuals was obtained by means of Partial Hausdorff Distance. Liu et al.\(^\text{22}\) experimented multi-views 2D face recognition performances of ICA and PCA, exploring the issues of subspace selection, algorithm comparison, and majority voting strategies. The second type of approaches is face recognition across pose, which needs a face alignment preprocessing stage to generate a novel frontal view image.

2. Previous Works

Estimating head pose of a subject observed by a camera is a very well studied problem in computer vision. Solutions adopted to tackle it can be subdivided into two main categories, as they make use of 2D or 3D information.

For the first type, they can be categorized in three different classes: feature-based, model-based and appearance-based approaches. In feature-based ones, a set of specific facial features, such as the eyes,\(^\text{23}\) nose, and mouth are used to estimate the head pose. In Hu et al.\(^\text{24}\) five facial features (the eye centers, the mouth centers and the nose) are localized within the detected face region. In model-based ones, some kind of head model is build upon a priori geometric knowledge. In Liu et al.\(^\text{25}\) and Ho et al.\(^\text{26}\) the human head is modeled with a 3D ellipsoid, so that any face image is a 2D projection of such a 3D ellipsoid at a certain pose. Appearance based ones, instead, concentrate on how the entire head image is modeled and learned from the training data. In Lablack et al.\(^\text{27}\) a template based method is presented as an image classification problem; the input head image is converted into a feature vector using SVD and Gabor wavelets; then, the feature vectors of different persons taken at the same pose are used to train two head pose classifiers, i.e., a RBF kernel SVM and a KNN. Introducing the temporal information to improve the head pose estimation leads to different algorithms, often referred to as head tracking. In Krinidis et al., a tracking technique, that utilizes a deformable surface model to approximate the image intensity, is used to track the face in the video sequence; subsequently, SVM\(^\text{28}\) or a RBF interpolation network\(^\text{29}\) is used to estimate the 3D head pose. Recent and reliable surveys in head pose estimation can be found in Murphy-Chutorian and Trivedi,\(^\text{30}\) and Ba and Odobez.\(^\text{31}\)

In a 3DFR context, correction of the relative position of human head points, prior to identification, is a mandatory task for the majority of methods proposed to date. Frequently, this operation is carried out manually by marking the position...
of known features (also known as fiducial points) and by adaptively compensating their location against a standard, average model. Despite the dominant trend, automatic registration has been attempted, and it may help to regroup existing methods into two main categories: registration and normalization methods. Registration ones are based on the comparison the subject under test against every single model in the database, or, in certain cases, against an ad hoc, average model. The classical Iterative Closest Point\(^{32}\) algorithm is still a very popular mean: Lu \textit{et al.}\(^{33,34}\) integrated the ICP-based rigid matching with the nonlinear Thin-Plate Spline (TPS) deformation; Irfanoglu \textit{et al.}\(^{35}\) also used ICP alignment to estimate landmarks automatically; Lonch \textit{et al.}\(^{36}\) performed two ICP stages for coarse and fine alignment; Gao \textit{et al.}\(^{37}\) adopted this algorithm for fine registration after that coarse alignment was achieved through nose-tip detection; Colbry and Stockman\(^{38}\) combined ICP with parametric geometry optimization, creating the so-called “Canonical Face Depth Maps”. Among other proposed means for registration it is worth to mention Kernel Correlation (Fabry \textit{et al.}\(^{39}\)) and a method by Pun \textit{et al.}\(^{40}\) in which the coarse step exploits a priori knowledge of the human face and facial features, and the fine step aligns the input data with the model stored in the database by the Partial Directed Hausdorff Distance.

Within normalization methods we include algorithms which really try to infer the head pose and orientation, by various means, and then correct it to obtain a frontal view. Natural mirror symmetry of the human face has been exploited for pose estimation in Hattori \textit{et al.}\(^{41}\). In Nanda and Fujimura\(^{42}\) a learning approach is implemented for single, 3D depth camera based, head pose estimation, in highly cluttered and complex real world environments. Many strategies have been employed to exploit 3D depth information to locate fiducial points, and then use these points to actually recover the head pose of the subject. For example, Li \textit{et al.}\(^{43}\) lean on the symmetry and features of human face to perform registration, then on profile and contours for recognition. Mian \textit{et al.}\(^{44}\) designed a robust method for nose-tip detection (also seen in Chew \textit{et al.}\(^{45}\)) and realized 2D/3D pose correction with Hotelling Transform. Mahoor and Abdel-Mottaleb\(^{46}\) obtained alignment based on three feature points, i.e., the inner corners of the two eyes and the tip of the nose, extracted through Gaussian curvature. Shin and Sohn\(^{47}\) performed automatic, face geometry based feature points extraction and used Error Compensated SVD for pose correction. Malassiotis and Strintzis’s\(^{48}\) work is also relevant as in which, a 3D face model is built using prior head-shoulder geometry knowledge together with probability assumptions; a Bayesian tracking framework is then used to track, in real-time, a subject’s movement when positioned in front of a range camera. Finally, particularly interesting are attempts by Blanz and Vetter\(^{49}\) to achieve pose estimation through a 3D morphable model, and then generate synthetic views of human faces with computer graphics, to account for variations in pose and illumination. He \textit{et al.}\(^{50}\) achieved pose correction through fiducial points and constructed synthetic views with RBFs and B-splines.
3. Model Reconstruction

The reasoning behind this first stage of preprocessing is two-fold: firstly, to recreate the worst operative condition or rather, that of a scanning device only able to capture clouds made of singular points than able to capture complete surfaces; and secondly, to deal with the unavoidable noise embedded in this type of tri-dimensional output. In order to create a mesh from single points, it is necessary to compute an approximation of the surface normals by relying on Dey and Sun’s \(^{51}\) algorithm (consisting in the smart application of Amenta and Bern’s \(^{52}\) polar balls method and enhanced by the accurate error model and noise-filtering thresholds). For precise surface estimation, Kazhdan, Bolitho and Hoppe’s Poisson Recon,\(^{53}\) is then used. The surface reconstruction process is based on an implicit 3D indicator function \(\chi\) (defined as 1 at points inside the model, and 0 at points outside). The main idea is that an integral relationship holds between the oriented points vector field \(V\), sampled from the surface of the model, and the indicator function. The problem of computing an implicit function \(\chi\) whose gradient best approximates \(V\), i.e., \(\min_{\chi}\|\nabla \chi - V\|\), can be restated as solving the following Poisson equation:

\[
\Delta \chi \equiv \nabla \cdot \nabla \chi = \nabla \cdot V, 
\]

i.e., compute a scalar function \(\chi\) whose Laplacian (divergence of gradient) equals the divergence of \(V\). The main advantage of this formulation is that the Poisson problem admits a hierarchy of locally supported functions, and therefore its solution reduces to a well-conditioned sparse linear system. Moreover, Poisson systems are well known for their resilience in the presence of imperfect data, which is usually unavoidable in a real-time context.

As the final step, the obtained 3D models (Fig. 1) are simplified through Garland’s vertex pairs contraction approach,\(^{54}\) (founded on the so-called Quadric Error Metrics and which is able to shrink a given 3D model to the desired complexity, in terms of polygon cardinality).

![Fig. 1. Original output mesh from 3D scanning device (left) and reconstructed model (right).](image-url)
The outcome results of this preprocessing stage are noise free, watertight and fairly simple human head models. In order to simplify subsequent analysis, their bounding box center is translated to axes origin. The reconstruction process is carried out in 5–15 seconds, depending on points cardinality and topology complexity (Fig. 2).

4. MSD and MPD

Normalizing the position of a human face implies the achievement of a perfect (or quasi-perfect) alignment of the 3D model of the head so that subject’s transformed model appears to be staring at the camera. This problem exposes the following six degrees of freedom: three angles of rotation (referred to as pitch, yaw and roll) and three measures of bounding box translation \((x_c, y_c, z_c)\), respectively relative to the three main axes \((X, Y, Z)\), as shown in Fig. 3.

The whole algorithm is conceived as a two-step optimization process.

In the first phase, we exploit the natural vertical symmetry of human faces; so gaining knowledge about the yaw and roll angles, along with the \(x_c\) translation
parameter. We have designed a simple measure of asymmetry between the two halves of a human head through the intersection of the 3D model with a hypothetical symmetry plane, and called it “Mean Symmetry Distance” (MSD), (Fig. 4). It is obtained by projecting a bundle of parallel lines which are perpendicular to the given plane, and intersecting them with the triangle mesh of the head. Then, for each line in the series, intersecting points closest to the plane are selected on either side (two points in total). The absolute difference between the distances of these points from the plane itself, is computed (where applicable: that is, if the projected line actually intersects each part of the head at least once).

Finally, the mean average of these absolute differences is calculated and taken as the symmetry error value (MSD) for the considered plane, as in:

\[
MSD_j = \frac{1}{n} \sum_{i=1}^{n} |dr_{ij} - dl_{ij}| ,
\]

where \( P_j \) is the considered plane, \( dr_{ij} \), \( dl_{ij} \) and are the distances between \( P_j \) and the intersection points \( pr_{ij} \) and \( pl_{ij} \), respectively, and \( MSD_j \) is the resulting, mean error measure for plane \( P_j \). The density of the line series may vary according to desired accuracy.

In the second phase, we lean on the natural vertical development of the central profile shape of human faces in order to gain clues about the pitch angle. To begin with, a sampling of the profile is extracted, resulting in a quasi-continuous point vector. This vector is then approximated through first-order equation parameters which have been adapted and in order to find the most suitable. In other words, we seek to find the line that best fits the shape of the profile, describing it by its explicit slope-intercept parameters. The error measure between the profile and its estimate, named Mean Profile Distance (MPD), is straightforwardly computed as the mean average of point-line distances between the given linear equation and the sampling vector (Fig. 5).
5. Implementation

At execution level, we looked for satisfactory ways to minimize the two error measures, MSD and MPD. In order to do so for MSD, the plane searching problem is inverted and solved with a multilevel genetic algorithm. Standard positions are fixed for the plane (coincident with \(YZ\)) and its relative, perpendicular line bundle (parallel to \(X\) axis). Hence, the test head model is rotated (around \(Y\) and \(Z\)) and translated (towards \(X\)) progressively within the GA to pull out yaw, roll and \(x_c\), whose floating point values constitute the genome.

The genetic algorithm is of Steady State type (with overlapping populations), implemented with GAlib.\(^{55}\) It divides the optimization process into three phases, with a gradual reduction of the search space and progressive stretching and thickening of the intersection grid. In the very first phase, short, sparse lines are used and error scores are divided by the number of actual intersections. In this way GA rewards transformations which maximize intersections cardinality; in particular, when located in the middle of face. This area in fact, contains much significant information about the frontal profile and is less likely to be affected by reconstruction artifacts. In subsequent investigations, longer and denser segments, coupled with proper MSD score handling, ensure the necessary final output accuracy. Validation of MSD measurements against outliers is carried out by filtering transformations for which the number of intersections fall below proper thresholds.

Initial boundaries of the variables to be optimized are set to \([-1.0; +1.0]\) radians for yaw and roll angles, and to \([-5.0; +5.0]\) units (approximately corresponding to 5 half millimeters) for \(x_c\). On the rim between one phase and the next, new boundaries are dynamically computed analyzing the current, score-ordered genetic population. The best individual score result is taken as reference and a number of other genomes that have score values within a varying upper limit range are also selected. The range is relative to the aforementioned reference and the current optimization
phase; however, at least ten genomes are selected (regardless of their score), to avoid premature convergence to local, suboptimal minima. Within this selected chromosome array, minimum and maximum values of each variable are detected and used as boundaries for the next phase. Evolution criteria is characterized by very high crossover (Blend) and mutation (Gaussian) rates, whilst termination is attained on a plain generation number basis. Results for this stage are noted in a pre-post representation in the following figure, with the $YZ$ symmetry plane highlighted in alpha-blended shaded color.

As for the second error measure, MPD, a greedy approach is employed. The idea is to center a proper line bundle at an empirically defined point (indirectly determining the intercept parameter); by progressively adjusting the slope with a trial and error procedure, a local MPD minimum is reached (Fig. 6). Initialization of the algorithm, lying in a feasible determination of line bundle center, is fulfilled in two ways: after extracting the profile sampling vector, the arithmetic mean and the min-max mean are computed for $y$ and $z$ coordinates ($x$ is null everywhere). For each of the two newly found points, slope angles in interval $[0; \pi]$ are sampled and scanned with a discrete step of $\Delta = \pi/4$ radians, while MPD is being computed for each of them. Angles associated with minimal error value are used to span a new, bounded line bundle with halved $\Delta$ value, as shown in Fig. 7.

Fig. 6. Head 3D model before (left) and after (right) MSD minimization process.

Fig. 7. Greedy MPD minimization process, with progressive halving of $\Delta$ step value.
The procedure is recursively performed until $\Delta$ falls under a certain threshold (0.001 radians) and final slope angles are elected as best pitch angle approximations.

Results of this algorithm, initialized with the above mentioned points, have been confronted with PCA, applied to the sampling vector. Being $P_{y1}$ and $P_{z1}$ the coordinates of the profile principal component:

$$Pitch = \tan^{-1} \left( \frac{P_{z1}}{P_{y1}} \right).$$

We discovered the most suitable output came from PCA and, partially, from min-max initialized MPD minimization, the results of which are, often, quite the same.

The last two degrees of freedom of the problem ($y_c, z_c$) are handled with simple a priori considerations about nose-tip position on angle-corrected models, which can be rightly assumed as the nearest point to the camera. Translating its $y$ and $z$ coordinates to axes origin is a simple and suitable solution.

6. Experiments

Experiments were carried out on the GavabDB 3D face database (collected by Moreno and Sanchez\textsuperscript{56} for Universidad Rey Juan Carlos, Madrid). Head point clouds were manually selected and data belonging to spurious parts of the scene (i.e., neck, shoulders, etc.) was cut out. Polygonal information (which actually builds the mesh of the model) was discarded during the reconstruction process.

Correct alignment can only be verified by an operator, from a qualitative point of view. Genetic algorithm settings were tuned conservatively in order to maximize the success rate in test circumstances (three attempts each with ten different subjects and seven poses each) for yaw and roll compensation. Perfect alignment was attained in more than 80% of the cases (162 times out of 195 total tests), whilst in the remaining cases (Quasi-perfect alignment), corrections within $[-0.02; +0.02]$ radians for yaw and roll, and $[-0.5; +0.5]$ for $x_c$ were necessary. The downside of this approach, however, is its extremely high computational cost: taking an average of 2 minutes’ time with an entry-level desktop configuration (AMD Athlon64 3200+, 1GB DDR Ram).
Table 1. Correctness rate of results yielded by the genetic algorithm for yaw and roll angles.

<table>
<thead>
<tr>
<th>GA result</th>
<th>Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perfect Alignment</td>
<td>83.08%</td>
</tr>
<tr>
<td>Quasi-perfect Alignment</td>
<td>16.92%</td>
</tr>
</tbody>
</table>

On the partially corrected model, PCA proved to provide the most stable and reasonable results (perfect pitch compensation in 100% of the cases). Min-max initialized MPD minimization challenges with PCA ca. 50% of the times, while arithmetic mean initialized MPD minimization results demonstrated it to be the worst behaved among the three. In 7.69% of the cases, both of the MPD minimum methods failed to supply correct results. Overall time requirements for each technique were below 1 sec.

Figure 9 contains random examples of the global compensation process output. In the “after” shots, symmetry plane is not visible due to the perfectly perpendicular position of the camera; please use the surrounding grid as a reference, instead.

Fig. 9. Some head 3D models before (up) and after (low) normalization.
Fig. 10. MSD values summary: thin vertical lines span MSD intervals, horizontal ticks mark their arithmetic mean while shaded bars represent Relative Standard Deviation.

The chart in Fig. 10 summarizes the MSD values recorded after yaw and roll normalization, coinciding with best genome’s final score after each run of the genetic algorithm. Minimum, maximum and mean MSD values are shown, together with relative standard deviation for each of the ten faces and considering the whole set of poses.

7. Conclusions and Future Works

We have designed a sturdy algorithm for the head pose normalization of 3D, scanned face models based on genetic algorithms, PCA and face geometry assumptions. Experiments conducted on the GavabDB database show a 100% success rate in correctly aligning 3D face models so that they “stare” at the camera (with absolutely perfect alignment in 83% of the cases). Future developments will surely imply a radical streamlining of the algorithm. At implementation level, a significant performance boost could be achieved by optimizing the algorithms to run entirely in the CPU caches, rather than in system memory. Furthermore, opportunities offered by highly parallel GPGPU architectures are still to be explored. From a theoretical point of view, we will firstly, have to replace GAs with simpler optimization techniques, like simulated annealing, hill climbing or other gradient descent approaches for example, as well as plan new, proper termination criteria. At the same time, space partitioning techniques will help reduce the computational complexity of the expensive ray-tracing like procedure underlying MSD calculus, by significant factors. Different reconstruction schemes should possibly be explored, in quest for the best trade-off between accuracy and speed (results from Zhou et al. should be taken into consideration). Attempts should also be made to integrate the methods herein into broader face analysis systems. Specifically, automatic recognition of fiducial points can provide significant initialization to optimization processes, reducing search space and execution time, while 3DFR technologies can effectively benefit from the achievements illustrated herein.
Appendix A. Genetic Algorithm Settings

<table>
<thead>
<tr>
<th>Table A.1. Genetic algorithm properties across the 3 phases.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>Generations</td>
</tr>
<tr>
<td>Population</td>
</tr>
<tr>
<td>Overlap amount</td>
</tr>
<tr>
<td>Grid thickness</td>
</tr>
<tr>
<td>MSD Score</td>
</tr>
<tr>
<td>Accuracy</td>
</tr>
<tr>
<td>Crossover probability</td>
</tr>
<tr>
<td>Mutation probability</td>
</tr>
<tr>
<td>Score validity threshold</td>
</tr>
<tr>
<td>Score selection threshold</td>
</tr>
</tbody>
</table>

References


48. S. Malassiotis and M. G. Strintzis, *Robust Real-time 3D Head Pose Estimation from Range Data* (Informatics & Telematics Institute of Thessaloniki, 2004).


55. http://lancet.mit.edu/ga/
