Face 3D Pose Estimation using a Generic 3D Face Model and Facial Features Extraction

Angelos Barmpoutis, Nikos Nikolaidis and Ioannis Pitas

AIIA Laboratory, Department of Informatics
Aristotle University of Thessaloniki, GR-54124 Thessaloniki, GREECE
Phone: +30 2310 996361, Fax: +30 2310 998453
E-mail: angelbar@csd.auth.gr, nikolaid@zeus.csd.auth.gr, pitas@zeus.csd.auth.gr
Web: http://poseidon.csd.auth.gr

Abstract. In this paper an algorithm that processes a video sequence for human face 3D pose estimation, is presented. The procedure that is followed is described briefly below. In the beginning, facial features are extracted for each frame. Afterwards, these are used in order to make an initial estimation of their position in 3D space. The results produced are optimized, either by taking into consideration anthropometric features, or by using a 3D model of the human face. In that way, the initial prediction is greatly improved and the resulting accuracy is more than satisfactory. Such techniques could process videos displaying news, journalists, actors or even people in general, or even be used in object-based techniques for video coding (eg. MPEG-4), in machine vision applications and in human computer interaction environments.

1 Introduction

Estimating the 3D pose or equivalently orientation of human face from still images or video sequences is important for object-based techniques for video coding (eg. MPEG-4), machine vision, face recognition, human-computer interaction, facial expression recognition etc. Thus, the problem of human face 3D pose estimation has attracted the interest of computer vision researchers in the last years. A variety of methods have been developed to approach this complex problem. Pose estimation using similarity measures was proposed in [7]. The use of Gabor wavelet networks for efficient head pose estimation has been proposed in [1]. A novel, efficient 3D pose estimation method from monocular video sequences is proposed in this paper. The method estimates the 3D pose by locating the eyes and the mouth on each frame and using a generic 3D model for refining the initial estimate.

According to the proposed technique the fist procedure that takes place is the localization of eyes and mouth in each frame of a video sequence, acquired by a calibrated camera. The facial feature extraction algorithm which was proposed in [2], has been used for this purpose. Having estimated the positions of these facial features on each frame, the method computes an initial estimate of the orientation in three dimensions of the human face. For this procedure we take into account anthropometric features, such as the fact that the eyes are equidistant from the mouth i.e., they form an isosceles triangle. The problem of face pose estimation essentially reduces into estimating the orientation of this triangle i.e., specifying the direction of its normal vector.

Since there are infinite points in three dimensions, which are projected at the same point on the projection plane (video frame), there is also an infinite number of triangles in three dimensions that have the same projection on the projection plane. Hence, there are infinite possible solutions for the pose estimation problem. Our main goal is to eliminate the number of the possible solutions, by introducing constraints derived from the characteristics and proportions of the human face. From all the possible solutions obtained, the one that is closer to our constraints is chosen as the initial estimate of face's 3D orientation.

This initial estimate is subsequently refined using a generic 3D face model. Refinement proceeds as follows: first a video frame where the initial pose estimation was successful is selected either manually or automatically and the image area corresponding to the depicted face is mapped as texture on the 3D face model. Then, for each video frame, the initial pose estimate is refined by searching for the orientation of the texture 3D face model which, when projected on the frame, yields the minimum error between the projected texture and the underlying image content. The results obtained, prove that the initial estimate is greatly improved.

In the following sections, the basic steps of the proposed face 3D pose estimation technique will be described in detail. Pose estimation results on head-and-shoulders video sequences will be also presented.

2 Facial Feature Extraction

The proposed 3D pose estimation technique is based on estimating in each video frame the position of the eyes and the mouth. This is essentially a two-step procedure that involves face region localization in the given frame and extraction of the facial features of interest (i.e. eyes and mouth) within the face region. Face detection and localization algorithms are based on texture, depth, shape, and color information. For example, face localization can take advantage of the fact that human faces are oval-shaped and have a distinct skin color. Another very attractive approach for face detection relies on multiresolution images by attempting to detect a facial region at a coarse resolution and, subsequently, to validate the outcome by detecting facial features at the next resolution.

Unfortunately, face localization and facial features extraction is not an easy task due to variations in luminance, facial expressions, view angles and potential existence of features like glasses, beard, etc. In the present paper, we used the feature extraction algorithm that was proposed in [2] and has shown to obtain very good results. However, other suitable facial feature extraction algorithms can be used for this purpose. According to this algorithm skin-like regions in an image are found by representing the image in the HSV (Hue-Saturation-Value) color space and by choosing proper thresholds for the values of the hue and saturation parameters. Considering the oval-like shape of a face, connected components of the skin-like regions are found, using a region-growing algorithm and an ellipse is fit to every connected component of nearly elliptical shape.

The extraction of eyes and mouth relies on the fact that they correspond to low intensity regions of the face. First morphological operations are applied on the image in order to enhance the dark regions and normalization of the orientation of the facial

region is performed. Afterwards, the x- and y- projections of the graylevel relief are computed. The y- projection of the graylevel relief is computed first and significant minima are determined based on the gradient of each minimum to its neighbor maxima. Then, the x- projection of the rows corresponding to the detected y-projection minima are computed, and minima analysis follows once again in order to find the horizontal positions of features. After constructing the lists of significant minima and maxima, candidate features are determined based on several metrics, e.g. the relative position inside the face, the significance of maximum between minima, the distance between minima, the similarity of graylevel values, and the ratio of distance between minima to head width.

3 Face 3D Pose Estimation

Having estimated the 2D coordinates of the eyes and mouth in each frame of the video sequence we proceed into estimating the 3D coordinates of these facial features. A point (x, y, z) in the three dimensional space is projected on the point (x_p, y_p) on the video frame (projection plane) and the following equations hold:

$$x_p = \frac{x}{\frac{z}{d}} \qquad y_p = \frac{y}{\frac{z}{d}} \tag{1}$$

In the previous equation, d is the distance between the projection center (i.e., the center of the camera lens) and the projection plane. d was evaluated by standard camera calibration procedures before proceeding to acquiring the video sequences.

There is an infinite number of points in the 3D space, which are projected at the same point on the projection plane. These points lie on a straight line that passes through this point (x_p, y_p) and the projection center. Thus, there are infinitely many triangles in the 3D space whose vertices are projected on the points corresponding to the centers of the eyes and mouth. Therefore the pose estimation problem has infinite solutions. In order to obtain a single pose estimate out of the infinite solutions, a number of constraints should be devised.

A basic constraint stems from the fact that the eyes are equidistant from the mouth, i.e. the three points form an isosceles triangle. Thus the search for a solution should be limited to isosceles triangles. Furthermore, since we are interested only in estimating the face orientation, (i.e. the orientation of the eyes-mouth triangle or equivalently the direction of its normal vector) and not the position of this triangle in the 3D space we can, without loss of generality, fix the position of one of the three points, namely the mouth in the 3D space. In other words, we assume that in the three-dimensional space, the mouth resides in an arbitrarily chosen point c in the corresponding straight line. By fixing the position of the mouth at point c, the locus of the eyes in the 3D space consists of the intersections of the two straight lines that correspond to the eyes with all concentric spheres centered at c, as can be observed in Figure 1. Hence, each of these spheres yields four solutions to the pose estimation problem, namely triangles $[p1, c, p2]$, $[p3, c, p4]$, $[p1, c, p4]$, $[p3, c, p2]$.

By observing Figure 1 one can easily notice that, as the radius of the concentric spheres increases, the four triangles resulting from each sphere correspond to faces

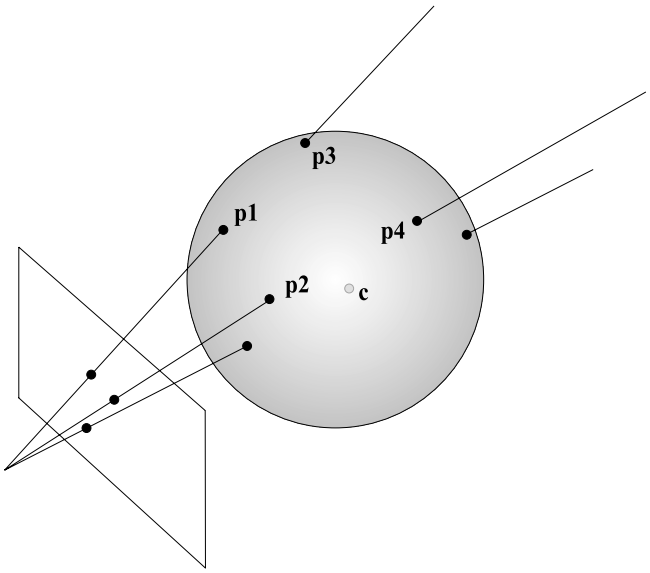

Fig. 1. Intersections between the sphere and the projection lines.

whose proportions are not "natural" (i.e., the eye-mouth distance is not in correct proportion to the eye-eye distance) and, furthermore, exhibit unnatural orientations (leaning forward or backward too much). Thus, the sphere with the smallest radius that intersects both lines corresponding to eyes is selected. Finally, among the four triangles defined by this sphere the one defined by points $[p3, c, p4]$ is selected as the initial pose estimate since the other three triangles correspond to less frequent orientations (the head is too much rotated around its vertical axis in case of triangles $[p1, c, p4]$, $[p3, c, p2]$ and too much forward leaning in case of triangle $[p1, c, p2]$). Furthermore, the selected solution corresponds to the triangle with the most "natural" proportions. It should be stressed here that the solution selection procedure described above gave the best results in practice. Furthermore, the solution obtained through this procedure need not be the optimal one, since a refinement procedure will follow (see next section).

Using the coordinates of these 3D points we can compute the rotation matrix that rotates the head from its "neutral" position (the eyes-mouth triangle is parallel to the projection plane and the vertical axis bisects the angle formed by the triangle's two equal sides) to the estimated position, considering the mouth point c as the fixed point of the rotation procedure.

4 Initial pose estimation refinement

The initial pose estimation procedure described above has been shown to yield satisfactory results. However this initial estimate might be inaccurate mainly due to erroneous estimates of the eyes and mouth position in the video frame. Thus, a refinement step that aims at improving the pose estimation accuracy has been introduced. A generic

three-dimensional geometric face model (triangular 3D mesh) has been used in this procedure. Using this model we can take advantage of our a-priori knowledge on the topological characteristics of the human face and introduce information about additional facial features (chin, nose, etc). Prior to its use in the refinement procedure that will be described below, the geometric face model is scaled to the correct proportions i.e. it is scaled so that its eyes and mouth consider with the three 3D points derived through the pose estimation procedure described in the previous section.

The refinement procedure is based in the following principle: if the 3D face model included texture information corresponding to the face depicted in the video sequence, then the accuracy of the estimated pose on an arbitrary frame could be judged by rotating and translating the 3D model to the estimated position, projecting the textured 3D model on the video frame and evaluating the mean squared (or mean absolute) error between the projected texture and the actual underlying image content, which from now on will be called projection error. In a similar manner, refinement of the pose estimate can be achieved as follows: starting from the initial pose estimate and using a certain error minimization strategy we rotate and translate the textured 3D model, in order to find the model orientation that yields the minimum projection error. The orientation of the model that yields the minimum error is adopted as the final pose estimate.

Obviously, in order to proceed in the refinement procedure described above, the face model should be enriched with texture information, i.e. a method for mapping the face region color information from an appropriately selected video frame on the (appropriately scaled) 3D model should be devised. A frame where the initial orientation estimation was sufficiently accurate is selected for this purpose. Prior to the texture mapping, the 3D model is rotated and translated according to this initial orientation estimate. Frame selection can be done either manually or automatically. Automatic frame selection, which was the procedure applied in this paper, is based on the observation that the initial pose estimation procedure described in the previous section is usually accurate in frames where the person is looking directly at the camera, i.e. it is at its "neutral" position. Thus, a frame where such a neutral position was returned by the estimation procedure is selected.

A projection error minimization strategy should be also selected. Since the refinement procedure applies translations and rotations on the 3D face model, the mean square error is a function of the translation and rotation parameters, i.e. it is a six-dimensional function. Searching this six-dimensional parameter space for the model position that yields the minimum error was performed using the three methods described below.

4.1 Random search

This error minimization strategy proceeds as follows: a number of random-valued six-dimensional candidate vectors, centered around the current pose estimate vector are examined and the one that yields the minimum projection error is selected as the current pose estimate for the next iteration. In its first iteration, the algorithm uses the pose estimate vector returned by the initial pose estimation procedure. The number of vectors examined in each iteration as well as the number of iterations that will be

performed are user-defined and control the complexity of the algorithm. Instead of performing a predefined number of iteration a termination criterion can be used. If the number of candidate vectors in each iteration is small (e.g. 100) the algorithm is very time efficient and can perform in real-time. The initial pose estimate is greatly improved.

4.2 Grid search

This minimization strategy proceeds like the random search strategy described above. However in each iteration the candidate vectors are placed on a regular grid around the current pose estimation vector. In the implementation used in this paper the grid consisted of 3^6 vectors resulting in a significantly increased execution time compared to random search. Obviously faster implementations can result by reducing the grid size.

4.3 Simulated annealing

The last minimization strategy that was used was the well-know simulated annealing algorithm which emulates the way liquids freeze and crystalize or the metals cool and anneal, reaching a minimum energy state [10]. The performance of simulated annealing depends on the appropriate selection of several elements of the method, e.g. the initial temperature, the annealing schedule and the random step selection procedure. For the parameters used in our implementation, the method performed slower than the random search strategy and did not yield significant improvements over the initial pose estimate.

5 Experimental results

The proposed algorithm was tested on several real head-and-shoulders video sequences with very good results. Two frames from different video sequences are shown in Figure 2. The initial pose estimates for these frames, presented using the estimated pose of the 3D model and the eyes-mouth triangle can be seen in the same figure. The derived textured 3D models (seen from different view angles) that were used in the refinement procedure, are also presented in the same figure. An example of the significant improvements introduced by the refinement procedure can be seen in the Figure 3. In most cases the refinement procedure corrects the initial pose estimation errors that result from inaccurate facial features extraction.

6 Acknowledgements

This study has been partially supported by the Commission of the European Communities, in the framework of the project IST-1999 20993 CARROUSO (Creating, Assessing and Rendering of High Quality Audio-Visual Environments in MPEG-4 context).

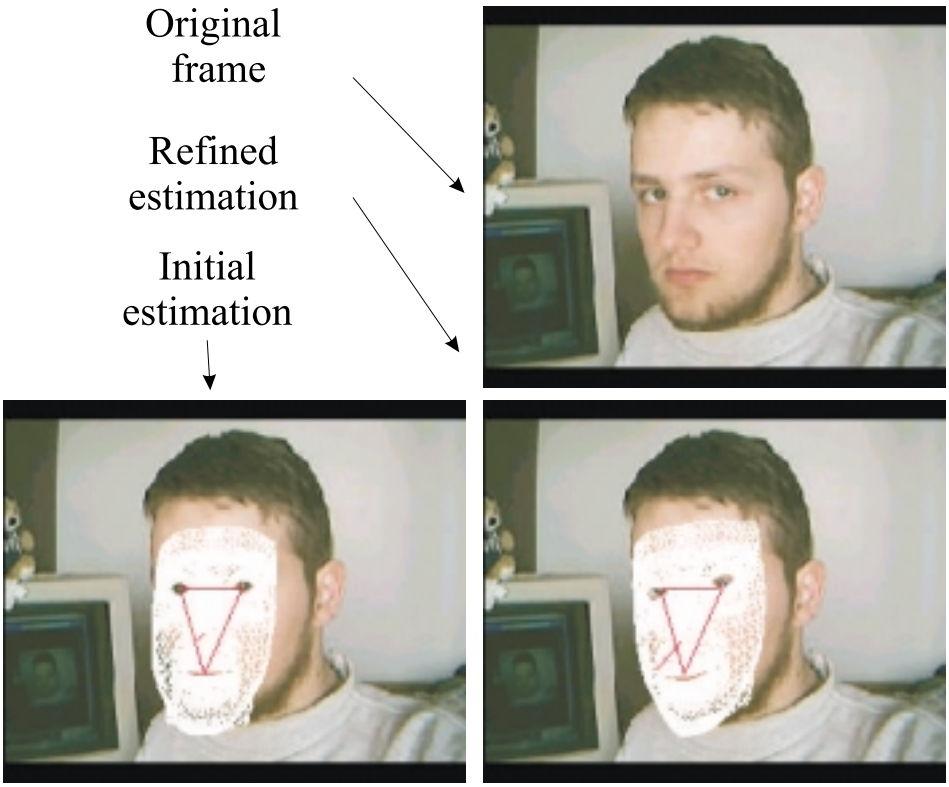

Fig. 2. Face 3D pose initial estimation examples and the derived textured face models.

References

1. Volker Krueger, Gerald Sommer: Gabor Wavelet Networks for Efficient Head Pose Estimation. Proceedings of British Mashine Vision Conference, Bristol UK,(2000).
2. Athanasios Nikolaidis and Ioannis Pitas: Facial feature extraction and pose determination. Pattern Recognition, Elsevier, (2000), vol.33, no.11, pp.1783-1791.
3. Edward S. Angel: Interactive Computer Graphics: A Top-Down Approach with Open-GL. Second Edition, 2000.
4. T. K. Leung, M. C. Burl, and P. Perona: Finding faces in Cluttured Scenes using Random Labeled Graph Matching. Proceedings of ICCV, (1995), pp.637-644.
5. L. G. Farkas: Anthropometry of the Head and Face. Raven Press, Second edition, New York.
6. Q. Chen, H. Wu, T. Fukumoto and M. Yachida: 3D head pose estimation without feature tracking. Proceedings of 3rd IEEE International conference on Automatic Face and Gesture Recognition, Nava, Japan, (1998).
7. J. Sherrah, Eng Jon Ong and Shaogang Gong: Estimation of Human Pose uning Similarity Measures. http://www.dcs.qmul.ac.uk/ sgg/pose/.
8. I. Shimizu, Z. Zhang, S. Akamatsu and K. Deguchi: Head pose determination from one image using a generic model. Proceedings of 3rd IEEE International conference on Automatic

Fig. 3. Initial pose estimation refinement example.

Face and Gesture Recognition, Nava, Japan, (1998).
9. K. Sobottka and I. Pitas: A novel method for automatic face segmentation, facial feature extraction and tracking. Image Communication, Elsevier, (1998), vol.12, no.3, pp.263-281.
10. Numerical Recipes in C++. Cambridge University Press, 2000.