ROBUST HEAD POSE ESTIMATION BY MACHINE LEARNING

Ce Wang and Michael Brandstein

Division of Engineering and Applied Sciences
Harvard University
Cambridge, MA 02138, USA
wangc,msb@hrl.harvard.edu

ABSTRACT

Support Vector Machines are applied for estimating the head orientation angle of talkers in a video environment. The procedure is capable of accurately evaluating head orientations over a complete 360 degree interval and has been designed to function as part of an existing real-time, multi-talker tracking system. By relying on a facial criterion that is easily extracted from video images acquired across a range of lighting and zooming conditions, the estimator is designed to be effective in practical situations such as those encountered in video conferencing or surveillance scenarios.

1. INTRODUCTION

A popular area of multi-media research is automated video conferencing in which a set of computer-controlled cameras capture the images of one or more individuals, adjusting for orientation, range, and source motion. The pose angle information is also useful for determining scene relevance and performing shot selection. In [1, 2] we proposed a hybrid face tracking system which made use of an acoustic-based localization algorithm for initial camera steering followed by a motion-based technique to develop body contours and detect facial regions. The result was a computationally efficient multi-source tracker.

The primary contribution of this paper is to introduce a Support Vector Machine (SVM) learning-based method for head pose angle estimation. Given the segmented head regions detected via the multi-talker tracker, the proposed algorithm extracts information relating to the hairline contour and learns an appropriate mapping between the resulting feature vector and the head orientation angle. This method is shown to be effective for facial orientation estimation in environments that provide less than ideal quality images.

2. FACIAL POSE ESTIMATION

A number of methods for facial orientation or head pose estimation have been proposed in the literature [3, 4, 5, 6]. The general approach involves estimating the positions of specific facial features in the image (typically the pupils, the corners points of eyes, the nose and mouth) and then fitting this data to some form of a head model. The accuracy and reliability of the feature extraction process plays an important role in the pose estimation results. In practice many of these methods still require manually selecting feature points, as well as assuming that near-frontal views and high-quality images are available. For the applications addressed by our work, such conditions are usually difficult to satisfy. Specific facial features are typically not clearly visible due to poor lighting and excessive source to camera depths. They may also be entirely unavailable when faces are oriented sufficiently far away from frontal views. Methods which rely on a detailed feature analysis followed by head model fitting fail to perform satisfactorily under these circumstances. We now propose a method for facial pose estimation which is both robust to the environmental conditions encountered and computationally simple enough for real-time application.

2.1. Feature Selection and Detection

In situations where most of the fine facial details, (e.g. eyes, lips, and mouth) are difficult to estimate reliably, we focus on a more overt facial property, namely the border between an individual's hair and skin, as the major feature for estimating head orientation. The hair-skin border is straightforward to estimate across a wide range of viewing angles and relatively insensitive to lighting conditions and poor image quality. As will be shown, the curve shape of the hairline is sufficient for estimating head orientation to a satisfactory degree and is a robust and consistent statistic across a variety of people and hairstyles. Figure 1 illustrates the hairline contour derived from three people in different orientations.

The hairline is extracted and parameterized into a simple feature vector. As shown in Figure 2, denoting the extent of the hairline along the horizontal axis as L, the hairline is evenly segmented into 6 pieces. The average vertical position of the contour segment for each piece, Y_i, is normalized

0-7803-6297-7/00/$10.00

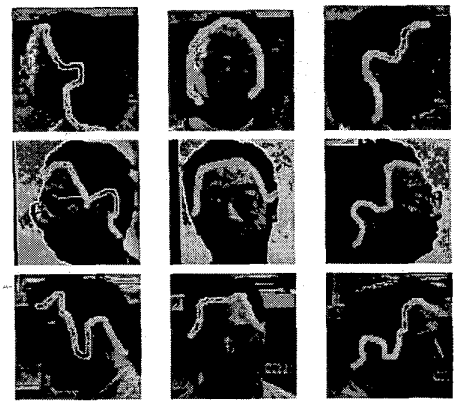


Figure 1. Hairline Patterns

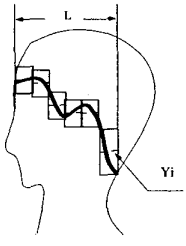


Figure 2. Parameterization of the Hairline Contour

to form the 6-point feature vector $\vec{y}$, by

$$y_i = \frac{Y_i - (Y_3 + Y_4)/2}{L}, i = 1, \ldots, 6$$

The purpose of this normalization is to make the feature vector invariant to the head size and relative frame position.

2.2. Learning and Estimation Procedures

The hairline contours shown in Figure 1 display specific patterns for various facial orientations. The patterns are quantized by the feature vector $\vec{y}$. The problem of orientation estimation becomes one of extracting the relationship between hairline patterns and head orientations. It may be formalized by finding the optimal mapping function $f : \vec{y} \rightarrow \theta, \theta \in [-\pi, +\pi]$, where θ is the facial orientation.

To increase the estimation accuracy, the training samples are first partitioned into three classes C_1, C_2 and C_3, based upon their coarse orientation angles. Figure 3 illustrates these classes which roughly correspond to a right, left, and rear facing head orientations. The frontal view is assigned a 0° orientation angle. The training data are assigned to one (or more) of three subsets: $S_1 = \{i|\theta_i \in [-120°, 0°]\}$, $S_2 = \{i|\theta_i \in [0°, 120°]\}$ and $S_3 = \{i|\theta_i \in [-180°, -120°] \bigcup [120°, 180°]\}$. To insure a continuous range

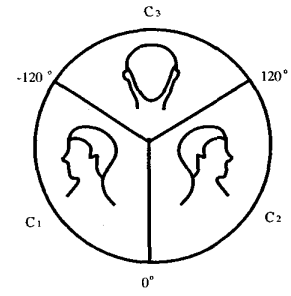


Figure 3. Subset classification.

of orientations all negative angles in S_3 are converted to positive value by adding 360°.

Within each class, the local mapping function $f_L i$ is obtained by training an SVM for regression [7]. The local mapping function for each class and its corresponding subset of training vectors, S_i, is given by:

$$\theta = f_{Li}(\vec{y}) = \vec{w}_i \cdot \vec{y} + \beta_i$$

It is obtained by solving the following minimization problem:

$$\min_{\vec{w}_i, \xi, \xi^*} \left[\frac{1}{2} ||\vec{w}_i||^2 + K_i \left(\sum_{j \in S_i} \xi_j + \sum_{j \in S_i} \xi_j^* \right) \right]$$

Subject to the constraints:

$$\begin{aligned} \theta_j - \vec{w}_i^T \vec{y}_j - \beta_i &\leq \epsilon_i + \xi_j \quad j \in S_i \\ \vec{w}_i^T \vec{y}_j + \beta_i - \theta_j &\leq \epsilon_i + \xi_j^* \quad j \in S_i \\ \xi_j &\geq 0 \quad j \in S_i \\ \xi_j^* &\geq 0 \quad j \in S_i \end{aligned}$$

where $||\vec{w}_i||^{-1}$, called the margin, is related to the lower bound on the distance between the points $\vec{y}_j$ and the hyperplane $(\vec{w}_i, \beta_i)$. K_i and ϵ_i are predefined parameters. The former controls the trade off between margin and the empirical risk which is proportional to $\sum \xi_j + \sum \xi_j^*$; the latter determines the deviation tolerance of training data from the regression. ξ and ξ^* describe the values outside of the tolerance limit ϵ. The idea underlying SVM for regression is to minimize the empirical risk as well as enlarging the margin at the same time. Details for finding the optimal $(\vec{w}_i, \beta_i)$ are available in [7]. This solution gives the training of local mapping functions.

2.3. Subset Classification

An observed feature vector $\vec{y}$ is first classified into one of the three subsets using a Bayesian Classifier, i.e. finding the class which maximizes the *a posteriori* probability given by:

$p(C_j|\vec{y}) = p(\vec{y}|C_j)p(C_j)$, $j = 1,2,3$. The *a priori* term $p(C_j)$ is calculated from the ratio of the subset cardinality, $|S_j|$, to the total number of training pairs, N, i.e.

$$p(C_j) = \frac{|S_j|}{N}$$

and class-conditional density is found from

$$p(\vec{y}|C_j) = \frac{1}{|S_j|} \sum_{i \in S_j} \Phi(\vec{y}, \vec{y}_i, \Sigma).$$

Without loss of generality, assume C_1 is the class which maximizes $p(C_j|\vec{y})$ and $C_k, k \in \{2,3\}$ is the class which is second most likely. If the likelihood associated with Class 1 is significantly greater than that of the C_k then the pose orientation is estimated directly from the class 1 projection. Specifically, given a predefined threshold, δ, if

$$|p(C_1|\vec{y}) - p(C_k|\vec{y})| > \delta$$

then the estimated orientation is :

$$\theta(\vec{y}) = \hat{\theta}_1 = \vec{w}_1^T \vec{y} + \beta_1$$

Otherwise the orientation angle is found using a weighted sum of the individual subset orientation angles according to:

$$\begin{aligned} \theta(\vec{y}) &= (\vec{w}_1^T \vec{y} + \beta_1) \left(\frac{p(C_1|\vec{y})}{p(C_1|\vec{y}) + p(C_k|\vec{y})} \right) \\ &+ (\vec{w}_k^T \vec{y} + \beta_k) \left(\frac{p(C_k|\vec{y})}{p(C_1|\vec{y}) + p(C_k|\vec{y})} \right) \end{aligned}$$

Note that the number of rough classes into which the continuous range of orientations are segmented is subject to the balance between the efficiency of the local prediction and the class separability. Increasing the number of classes improves prediction accuracy of the linear projection, but at the cost of decreased subset classification performance. The use of three subset classes was found to provide a reasonable compromise between these effects.

The overall procedure is outlined in Figure 4.

2.4. Performance Demonstration

Supervised training for the proposed scheme was achieved using a total of 60 hand labeled vector-orientation pairs derived from video sequences of the three individuals shown in Figure 1 across the full range of potential orientation angles. The individuals were selected to provide a variety of hairstyles. Figure 5 illustrates the pose estimation results achieved for sample images of the video stream for a fourth subject. In each picture a single image frame is shown along with a clock indicating the estimated facial orientation (6 o'clock corresponds to fully forward-facing, 12 o'clock is fully rear-facing) and a plot illustrating the segmented head region along with the detected hair and skin blobs used to extract the feature vector. A video version of this result is available from the webpage: *himmel.hrl.harvard.edu*.

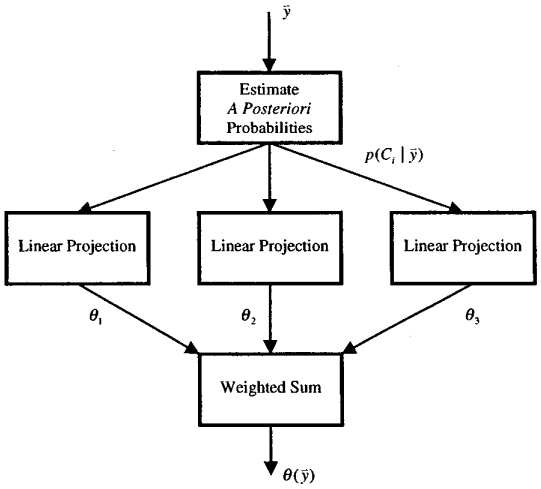


Figure 4. Pose Estimation Outline.

The pose estimation results are very accurate across the complete range of possible orientation angles despite the presence of object shadows and a complex image background. This illustrates the appropriateness of the extracted image features and the effectiveness of the proposed pose learning and estimation procedures. Our experiments indicate that the hairline contour can be reliably estimated from facial images encompassing a range of lighting conditions and source-camera distances using only relative intensity statistics, color disparity, and region connectedness. In situations where the skin or hair regions are partially misidentified, the pose estimator still provides reasonable results. Figure 6 illustrates an example where the head has been incorrectly segmented from the background and the hair blob takes on a peculiar shape. In this case the orientation is correctly estimated.

Besides accurately estimating orientation angles appropriate for their specific class, the local linear projections extrapolate well to pose angles outside their defined range. In Figure 7 a pair of images which would fall in the Class 2 angle region are estimated entirely from the projection associated with Class 1, i.e. $\theta(\vec{y}) = \hat{\theta}_1$. In both instances the estimated pose angles are very reasonable.

3. CONCLUSIONS AND FUTURE WORK

This paper has presented our work with pose estimation as applied to a face tracking system. Video clips of the results demonstrated here as well as the overall tracking system are

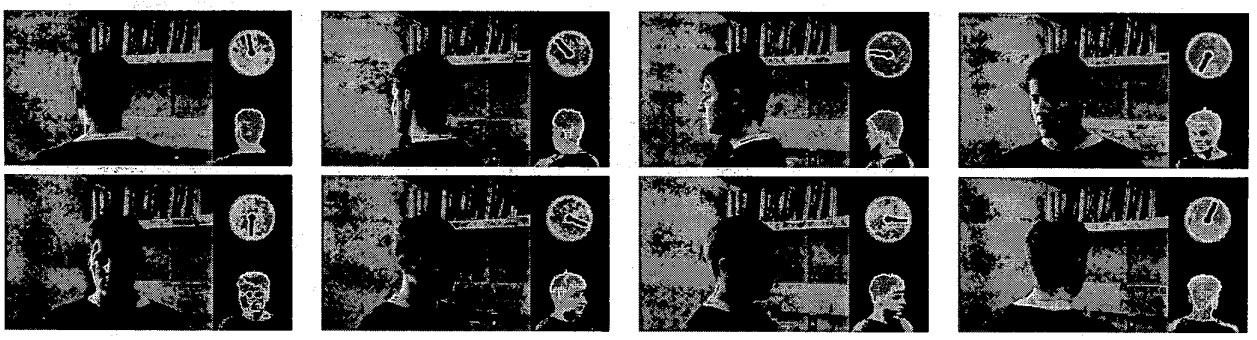
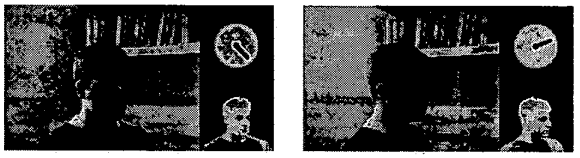


Figure 5. Demonstration of Face Orientation Estimation Results

Figure 6. Pose Estimate for a Misidentified Hairline Contour.



(A) (B)

Figure 7. Estimation of a Class 2 Pose Angle using a Class 1 Projection

available from the webpage: http://himmel.hrl.harvard.edu. The pose estimation procedure introduced is capable of accurately evaluating head orientations over a complete 360° interval. The algorithms are designed to be simple enough for real-time applications and, by relying on facial criteria that are easily extracted from images across a range of lighting and zooming conditions, to be effective in practical environments such as those encountered in video conferencing or surveillance scenarios.

A number of avenues are available for improving the capability and functionality of the proposed method. The training and testing images will be expanded to a more substantial sample size encompassing a larger range of people and hairstyles. An obvious point is that the algorithm relies on a criterion which many individuals may not possess, namely a hairline. To address this issue we plan to extend the pose estimation scheme to include additional facial features such as the rough locations of the eye and mouth regions. Ideally, the estimator will incorporate a number of facial cues and will take into consideration their reliability when calculating a final result. Finally, in this work, only the pan rotation of the head is considered. The learning-based method employed here may be adapted to evaluate other orientation features such as tilt.

4. REFERENCES

[1] Ce Wang and Michael S. Brandstein. A hybrid real-time face tracking system. In *ICASSP'98*, volume 6, pages 3737–3741, Seattle, Washington, May 12-15 1998.

[2] Ce Wang and Michael S. Brandstein. Multi-source face tracking with audio and visual data. In *Proceedings of IEEE 3rd Workshop on Multimedia Signal Processing*, pages 169–174, Copenhagen, Denmark, September 13-15 1999.

[3] Ricardo Lopez and Thomas S. Huang. Head pose computation for very low bit-rate video coding. In *6th International Conference on Computer Analysis of Images and Patterns*, pages 440–447. Springer-Verlag Berlin Heidelberg, 1995.

[4] Thomas Maurer. Estimation of face position and pose with labeled graphs. In *BMVC*, 1996.

[5] N. Kruger, M. Potzsch, and C. Malsburg. Determination of face position and pose with a learned representation based on labeled graphs. *Image and Vision Computing*, 15(8):665–673, August 1997.

[6] I. Shimizu, Z. Zhang, S. Akamatsu, and K. Deguchi. Head pose determination from one image using a generic model. In *3rd IEEE International Conference On Automatic Face and Gesture Recognition*, pages 100–105, Nara, Japan, April 1998.

[7] V.N. Vapnik. *Statistical Learning Theory*. Wiley, 1998.