Real-time Driver Affect Analysis and Tele-viewing System[i]

Joel C. McCall, Satya P. Mallick, and Mohan M. Trivedi
Computer Vision and Robotics Research Laboratory,
Department of Electrical and Computer Engineering,
University of California, San Diego.

Abstract

This paper deals with the development of novel sensory systems and interfaces to enhance the safety of a driver who may be using a cell phone. It describes a system for real time affect analysis and Tele-viewing in order to bring the context of the driver to remote users. The system is divided into two modules. The affect recognition module recognizes six emotional states of the driver. The face-modeling module generates a 3D model of the driver's face using two images and synthesizes the emotions on it. Depending on bandwidth, either a raw video, 3D warped model, or an iconic representation is sent to the remote user.

1. Introduction

Cell phone use while driving is recognized as a significant distracter for a driver [8]. This research deals with the development of novel sensory systems and interfaces to enhance the safety of a driver who may be using a cell phone. The work is based on the premise that a safer mode of communication between a remote caller and the driver can be achieved by providing the remote caller with a visual feedback of the drivers affective state as well as an indication of the driving conditions [9] in real-time. The flow chart for this system is shown in figure 1.

The real-time facial affect analysis is performed using a single camera viewing the driver. By tracking facial landmarks, a feature

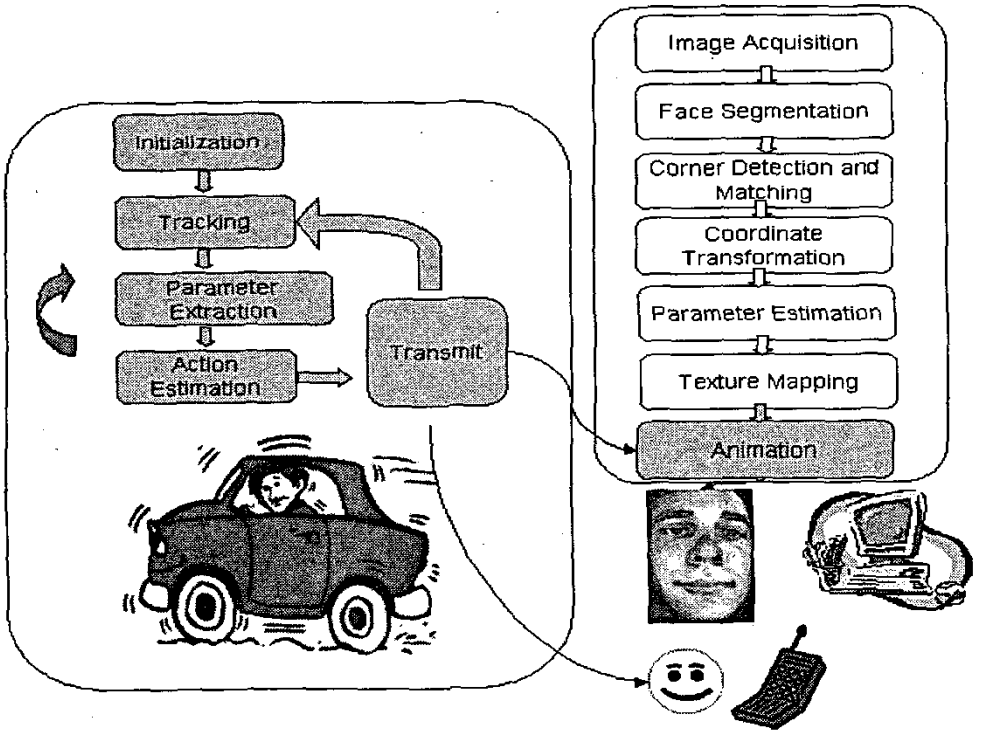

Figure 1: Conceptualization and algorithmic modules for a driver affect analysis and tele-viewing system

0-7803-7848-2/03/$17.00 ©2003 IEEE

vector containing the nonlinear warping of the drivers face can be generated. This feature vector can be used as an input to a classification system for determining the drivers facial expression from one six universal facial expressions. These expressions are happiness, fear, sadness, surprise, anger, and disgust. The emotion parameter and warping parameters are then sent to the Tele-viewing system that creates a representation of the drivers face.

The basic idea of the Tele-viewing system is to represent the state of the driver's face, using only a few parameters. The emotion specified by the parameters can then be reconstructed at the receiver end. The bandwidth bottleneck is thus overcome. However, to represent the emotions realistically, it is not sufficient to send the information about the driver's state; good reconstruction and display of the emotions is of extreme importance for Virtual Tele-presence. A stereo setup is used to capture two images of the driver's face with no emotions (a neutral face). A 3D model of the face is constructed using the two images. The 3D model is a triangular mesh model with 185 points. The mesh model is automatically texture mapped using one of the images of the driver's face. The face texture and the 185 points are transmitted to represent a neutral state (no emotions). The six emotions described can be synthesized by a transformation of the coordinates of this mesh model.

2. Real-Time Facial Affect Analysis

The following section describes the subsystem developed for facial affect analysis using Bayesian estimation and Hidden Markov Models. The subsystem is organized into four main components and is designed to run in real-time inside a vehicle. The initialization routine takes user input to determine the neutral feature templates used in the tracking mechanism. Once initialized, the program loops through tracking, parameter extraction, and expression recognition.

2.2 Feature Tracking and Feature Vector Calculation

After initialization, feature tracking is performed in real time by use of an affine transformation model described by (7) and (8),

$$u = a_0 + a_1 x + a_2 y \quad (7)$$
$$v = a_3 + a_4 x + a_5 y \quad (8)$$

The value of u represents the difference in the x direction and the value of v represents the difference in the y direction. These tracked points are then used to calculate the thin-plate spline warping of the face. For details on this method of affect estimation, please refer to [6] for details of this method of affect analysis.

2.3 Hidden Markov Models for Expression Estimation

The computed parameters are used to generate an estimate of the facial expression. This is done using a Hidden Markov Model of the expression states. The outputs from the hidden states are determined from the metrics extracted in the previous step. Using the HMM, a maximum a posteriori estimate of the expression state is computed. A general tutorial on HMMs can be found in [7].

The seven-state HMM used in this system contains states corresponding to the expressions Neutral, Happy, Sad, Anger, Fear, Disgust, and Surprise. The state transition matrix is constructed with probabilities that favor staying in the current state. The HMM is also useful in preventing jumps between dissimilar states.

The final estimation step is done using a Bayesian maximum a posteriori estimation method. Given a previous state X_p and output Y, the current state X is estimated by the following equation:

$$X = \arg\max P(Y|X_i)P(X_i|X_p) \quad (9)$$

The probability mass function P(X) is taken from the HMM state transition matrix. Specifically $P(X_i|X_p)$ represents the probability of transitioning from state X_p to state X_i. The probability density function (pdf) $P(Y|X_i)$ is computed by training the system from known facial expressions. This training is done by acquiring data representing known facial expressions and using a maximum likelihood estimation to determine the best fit Gaussian pdf.

The output of the face-affect analysis system along with the landmark points warping vector is then sent to the Tele-viewing subsystem where a Model based reconstruction of the subject current expression is generated.

3. Parametric Face Modeling

Recovery of detailed depth information using dense stereo is extremely noisy and, hence, not suitable for face modeling [1]. We propose a sparse stereo algorithm for modeling of human

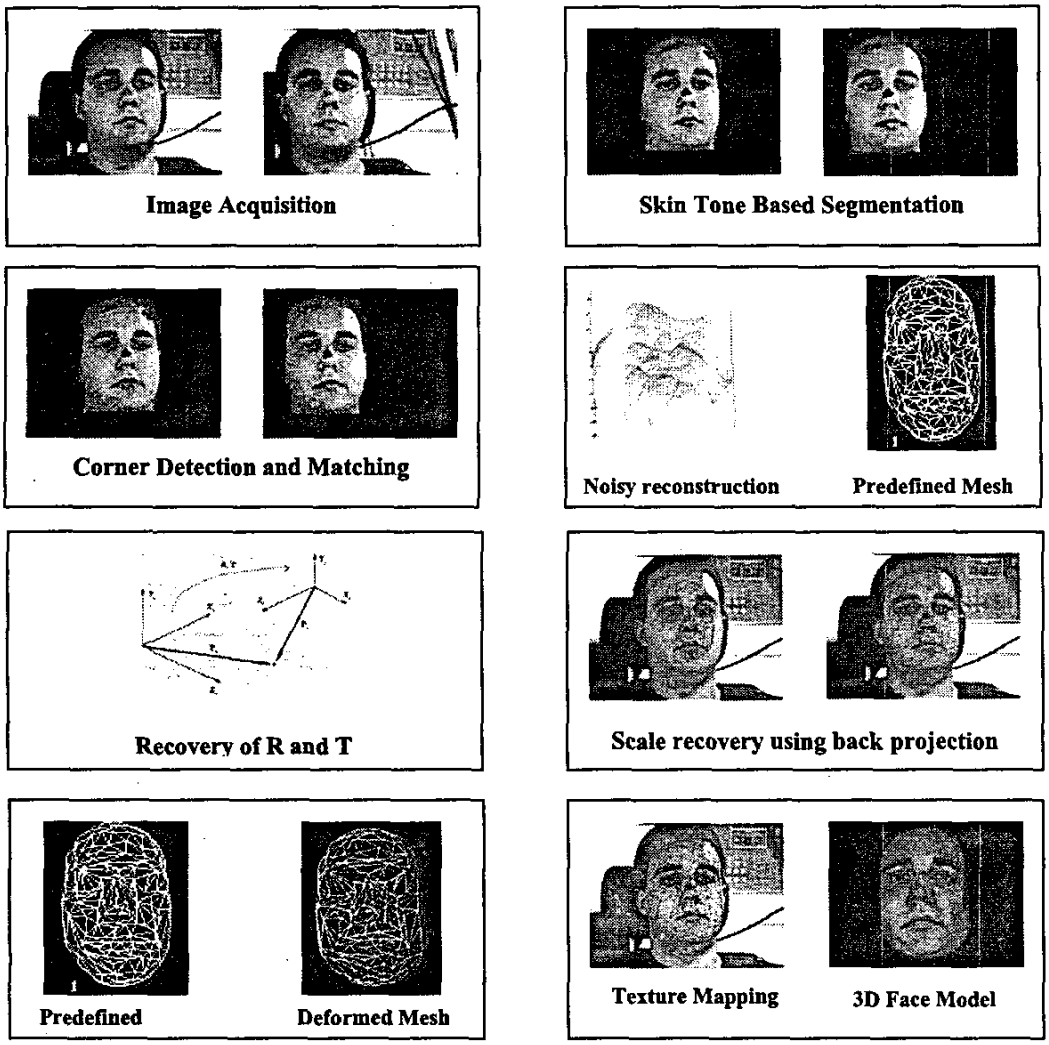

Figure 2: Stepwise depiction of the parametric face modeling algorithm

faces using a calibrated stereo setup. Sparse stereo data cannot be totally trusted for face modeling, because even a small error in reconstruction leads to visually unrealistic results. Hence, our approach is to deform a predefined parametric face model to fit the stereo data obtained using sparse stereo matching. Different parameters define the shapes and sizes of the different components of the face. The deformed mesh is automatically

Different parameters define the shapes and sizes of the different components of the face. The deformed mesh is automatically texture mapped with the face texture of one of the images.

3.1 Image Acquisition and Segmentation

The images of size 640x480 are acquired using a calibrated stereo rig. The face is segmented out from any arbitrary background using skin tone based segmentation [4]. Skin tone based segmentation is extremely fast, robust and doesn't require any manual tweaking for a wide range of skin tones.

3.2 Corner Detection and Matching

Harris corner detector is used to detect corners on both the images. Several methods were tried for solving correspondences among the detected corners. We currently use Zhang's relaxation based method [3] and a Singular Value Decomposition (SVD) based method to solve for correspondences [2]. The 3d coordinates of the corresponding points are calculated using the calibration information of the cameras. An attempt to reconstruct the face, using the depth values of the corresponding points would result

in an extremely noisy reconstruction. Hence, such an approach is not taken.

3.3 Coordinate Transformation

We dealing with two coordinate systems: 1) the coordinate system in which the predefined mesh resides and 2) the coordinate system in which 3d coordinates of the corresponding points reside. We need to transform one coordinate system to another. Five corresponding points on the face are hand-clicked in the two images and the depths of the five points are recovered using the calibration information of the cameras. The five points are: the two ends of the lips, the tip of the nose, and the inner ends of the eye. The coordinates of these five points in the predefined mesh are known. Hence, we can recover the rotation (R), translation (T), and scale (s) relating the two coordinate systems [5]. The scale recovered using this method is not reliable. Using the R and T values, the coordinates of the predefined mesh are projected onto the image plane of the left camera. The amount of scaling required for the projected mesh to touch the bounding box containing the segmented face gives us the scale (s).

3.4 Solving 3D correspondences using Hungarian algorithm

All the corresponding points are transformed to the coordinate system in which the predefined mesh resides. We call this set of points P. The predefined mesh is made of 185 points. We call this set of points P_{mesh}. For each point in P_{mesh} we find a point in P, which corresponds most closely with it using Hungarian linear partitioning algorithm. We call this set of points n_{mesh}. n_{mesh} is mesh that is obtained using actual data and P_{mesh} is the predefined mesh. The next step is to change the parameters of P_{mesh} so that it matches n_{mesh}.

3.5 Parameter Estimation

The P_{mesh} can be deformed by changing a few parameters. Each parameter defines a particular attribute of the face; like the shape and size of the nose, placement of the eyebrows, size of forehead etc. Each parameter has an upper bound and a lower bound. This constraint ensures that we do not over fit the noisy n_{mesh}, which can result in an unrealistic face model. Let,

$\hat{n}_{mesh}$ = Estimate of n_{mesh} by deformation of P_{mesh}.

$$\hat{n}_{mesh} = P_{mesh} + \sum_{i=1}^{n} \alpha_i d_i$$

$$-1 \le \alpha_i \le 1$$

$$C = \left| n_{mesh} - \hat{n}_{mesh} \right|^2$$

Where C is the cost function we seek to reduce. α_i's represent the amount of deformation, and d_i's are the deformations added to P_{mesh}. The Cost function C can be broken down into smaller independent cost

$$C = \sum C_i$$

functions, each of which is much easier to solve. For example, we can change the size of the nose independent of the size of the forehead. Hence, we have the new cost function,

3.6 Texture Mapping

$\hat{n}_{mesh}$ is projected onto the left image and the texture enclosed within a triangle formed by the vertices of the projected mesh is mapped onto the corresponding 3d triangle in $\hat{n}_{mesh}$.

4. Affect Estimation Results

The following still images taken from a processed video show the results of the system after it has been trained. The bars indicate the relative probabilities associated with each of the expressions and the face icon indicates the final result of the affect estimation. The points on the face are those that are being tracked from frame to frame for the extraction of feature vectors.

Tests were performed in both a lab setting as well as within a moving vehicle. The lab environment consisted of stable lighting conditions and slight head movements. The vehicular environment consisted of a car moving at a speed of 30-40 mph on an afternoon with slightly cloudy sky. There was a considerable amount of jerking movement due to the vehicle moving on the uneven road surface. In both cases the estimation accuracy was seen to be above 90%; however, the tracking accuracy was reduced in the vehicular setting because of fluctuations in lighting as well as the jerking of the vehicle. This limitation in tracking can be overcome by adding more complex model based filtering as well as preprocessing to eliminate

Figure 3: Various expressions (Happiness, Sadness, Fear, Anger, and Disgust) performed inside a car and the resultant output expression to be displayed remotely.

some of the effects of lighting changes. Figures 3 and 4 demonstrate the results of the tracking and expression estimation.

Figure 4. A sample of expression recognition while the subject is using a cellular phone.

5. Face Animation Results

Emotions can be synthesized by a non-linear transformation of the vertices of the mesh. In our implementation, we first generated six meshes corresponding to the six emotional states, by manually moving the vertices of the predefined mesh. We calculate the percentage deformation of each vertex of the predefined mesh required to generate a particular emotion. The same emotion can be generated on the new mesh by moving each vertex of the mesh so as to bring about the same percentage deformation. Intermediate emotions can be generated by interpolating the coordinates of each vertex between two emotional states.

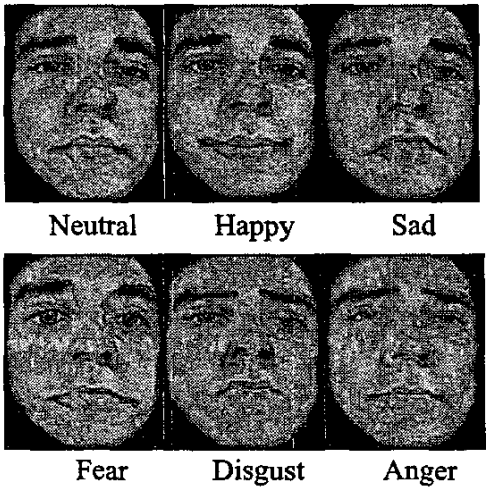

Figure 5. Face Animation Results

5.1 Smoothing using splines

In both face modeling and animation we are deforming a smooth mesh. The resulting deformation may sometimes result in a faceted appearance which can look quiet unrealistic. To overcome this problem, we use splines to smoothen the deformed mesh. In our implementation, we are using the in-built spline based smoothing in the VRML browser called "Cortona" by Parallel Graphics. The results are shown in Figure 6. There is a noticeable improvement near the nose, the cheekbones and the region between the eyebrows.

Figure 6. Result of smoothing using splines

6. Conclusion

In this paper we have presented our efforts towards the development of driver affect capture, analysis, and tele-viewing. The research is currently being integrated with other efforts associated with real-time vehicle surround analysis [9]. After such integration, systematic and careful human factors experiments will be preformed to evaluate the efficacy and utility of these modules.

References:

[1] Z. Liu, Z. Zhang, C. Jacobs, and M. Cohen. *"Rapid modeling of animated faces from video"*, *Journal of Visualization and Compute Animation*, 12(4):227-240, 2001.

[2] M. Pilu "A *direct method for Stereo Correspondence based on Singular Value Decomposition*", IEEE International Conference of Computer Vision and Pattern Recognition, Puerto Rico, June 1997

[3] Z. Zhang, R. Deriche, O. Faugeras and Q.-T. Luong. ``A Robust Technique for Matching Two Uncalibrated Images Through the Recovery of the Unknown Epipolar Geometry", *Artificial Intelligence Journal*, Vol.78, pages 87-119, October 1995.

[4] J. Yang, W. Lu, and A. Waibel. *"Skin-color modeling and adaptation"*. In Proceedings of ACCV, Hong Kong, volume 2, pages 687-694, 1998.

[5] B. K. Horn. *"Closed-form Solution of Absolute Orientation using Unit Quaternions"*. Journal of the Optical Society A, 4(4):629–642, Apr. 1987.

[6] McCall, J. and Trivedi, M, "Real-Time Facial Affect Analysis Using Thin-Plate Splines," Submitted to *International Conference Computer Vision 2003*.

[7] L. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proc. IEEE*, Vol. 77, No. 2, pp. 257-286, Feb. 1989.

[8] Donald A. Redelmeier, Robert J. Tibshirani, "Association between Cellular-Telephone Calls and Motor Vehicle Collisions," *The New England Journal of Medicine* -- February 13, 1997 -- Vol. 336, No. 7 (http://www.nejm.org/content/1997/0336/0007/0453.asp

[9] Tarak Gandhi, Ofer Achler, and Mohan Trivedi, "Vehicle surround analysis for intelligent driver support system using omni-directional video".

[10] Bassili, J.N. 1979. Emotion recognition: The role of facial movement and the relative importance of upper and lower areas of the face. *Journal of Personality and Social Psychology*, 37:2049-2059.

[11] Black, M. J. and Yacoob, Y. 1997. Recognizing facial expressions in image sequences using local parameterized models of image motion. *International Journal of Computer Vision*, 25(1):23-48.

[12] De la Torre, F., Yacoob, Y., and Davis, L. 2000. A probabilistic framework for rigid and non-rigid appearance based tracking and recognition. *International Conference on Automatic Face and Gesture Recognition*, (FG2000)

[13] Ekman, P. 1992. Facial expressions of emotion: An old controversy and new findings. *Philosophical Transactions of the Royal Society of London*, B(335):63-69.

[14] Essa, I. A. and Pentland, A. 1995, Facial Expression Recognition using a Dynamic Model and Motion Energy. *International Conference on Computer Vision*, '95, Cambridge, MA

[15] Lien, J. J., Kanade, T., Cohn, J. F., Li, C. 1999, Detection, Tracking, and Classification of Action Units in Facial Expression. *Journal of Robotics and Autonomous Systems*, July 28/August 21, 1999

[16] P. Ekman and W. V. Friesen. The Facial Action Coding System: A Technique for Measurement of Facial Movement. Consulting Psychologists Press, San Francisco, CA, 1978.

[i] *This research is sponsored by UC Discovery Grants and the Daimler Chrysler Corporation. The authors would also like to thank their colleagues in the CVRR Laboratory for valuable assistance and cooperation.*