Int. Conference on Pattern Recognition, Cambridge, UK, August 2004

Pose Invariant Affect Analysis using Thin-Plate Splines

Abstract

This paper introduces a method for pose-invariant facial affect analysis and a real-time system for facial affect analysis using this method. The method is centered on developing a feature vector that is more robust to rigid body movements while retaining information important to facial affect analysis. This feature vector is produced using thin-plate splines to extract affine transformations independently from nonlinear transformations quickly and efficiently. The affine portion can be used to describe the rigid body motion because planar motions in a perspective projection can be approximated by an affine transformation. Removing the affine portion and using the nonlinear portion of the thin-plate spline warping provides information on the nonlinear motion caused by facial affects.

The real-time system developed using this method is composed of three main components: facial landmark tracking, feature vector extraction, and affect classification. The system processes streaming video in real-time. Testing was performed to examine the invariance to rotation as well as subject independence of the system. Finally its application in real-world environments is discussed.

1 Introduction

Analysis of facial expressions by machine vision systems is an important research area for many applications. Applications ranging from user interfaces to intelligent vehicles and spaces can be greatly enhanced with the incorporation of expression analysis [1]. Although it has been actively researched, there are still many aspects of the field that are open research areas.

The difficulties in developing a facial expression recognition system lie in several different areas. These areas include difficulties imposed by lighting conditions, variations in the expressions between persons, and head pose and head movement [2].

Figure 1 Images showing the variations in lighting, head position and expression for anger.

Lighting conditions can also pose problems. Because of the complex face surface, small head movements or expressions can greatly change the way the face is lit and change shadows. This creates problems by creating shadows that move and obscure important features useful in identifying facial expressions. Algorithms that use optical flow might get false motion from these shadows. Creating robustness to head movements and head pose is also very desirable in facial expression systems. Head movement can pose big problems for algorithms that rely on feature vectors such as the principal components of the image or the image motion. Using many of these techniques requires training the head motion into the classifier rather than preprocessing the data so that only features which contain information on non-rigid motion are fed into the classifier. It would be very beneficial to have a feature vector that is invariant to such movements. Figure 1 shows the kind of variations that are present in facial expression analysis.

It is this motivation that leads to the development of the thin-plate spline feature vector and an implementation of a pose-invariant real-time machine vision system for analyzing facial expressions. By using the nonlinear portion of the thin-plate spline warping, we generate a rotation and translation invariant feature vector, thereby separating the rigid and non-rigid facial motion. The key components of our real-time system are facial landmark tracking, feature vector calculation with thin-plate splines, and classification. The rotational and translational independence helps reduce the complexity and amount of training required for the classifier.

This novel application of thin-plate splines provides a fast, efficient and robust way of estimating facial affects. We show that the feature vector generated by the thin-plate splines encompasses the statistics necessary to perform facial affect analysis while providing invariance to rotation and translations of the subjects face, thereby reducing the necessary complexity of the classifier. Invariance to lighting is dependent on the facial landmark tracker, and is offloaded from the feature vector extractor as a preprocessing step.

1.1 Related Work

Psychologists have developed an understanding for certain universal facial expressions. Specifically, it has been found that there are at least six universal facial expressions [3]. Others have discovered visual cues for determining facial expressions [4]. The non-rigid motion of specific facial features characterizes each of the six distinct facial expressions. These non-rigid motions can be categorized into individual "action units". Ekman and Friesen proposed a widely used Facial Action Coding System or FACS [5] that we will use to describe specific facial movements.

Many researchers have shown the significance of motion energy in facial affect analysis by analyzing the optical flow of image sequences [6,7,8]. However, these methods often break down when rigid body motion, due to head movement, is present. Therefore, the non-rigid body motion must be separated from the rigid body motion in order to provide robust facial affect recognition.

Others have attempted to solve this problem of separating the rigid motion from the non-rigid motion by using model-based estimation [7]. Unfortunately many of these techniques are too slow to be used in a real-time system. Also, a closed form solution is much more desirable than other iterative methods such as Generalized Procrustes Algorithms [9]. Thin-plate splines, however, can provide a separation of the nonlinear motion from the affine motion in a one-step closed form solution [10].

Non-rigid feature tracking has been explored in a variety of ways ranging from methods based purely on optical flow and probabilities, to methods which construct detailed facial models and perturb them according to facial landmark movement. Black and Yacoob have shown work that parameterizes various facial feature motions with affine and similar transformation models [11].

Another approach taken to this problem is to input more complex feature vectors into classification systems. Systems developed using Graph Matching [12], Neural Networks [13], and Support Vector Machines [14] have been shown to be effective, but require more complex classification schemes.

Parameter estimation has been examined in many different areas of research. Applications of these concepts to facial feature estimation are well demonstrated by various research groups [14,15,21]. Classification techniques involving HMMs [21], mixture densities, likelihood functions, principle component analysis, support vector machines, and independent component analysis have shown to be useful for these problems.

The unsolved problems of algorithmic efficiency and robustness to movements and environmental conditions motivated the development of the feature vector extraction and system described in this paper. By speeding up the point tracking algorithm to run in real time and introducing thin-plate splines to parameterize facial motion, we have developed a real-time system that demonstrates robustness to movement.

2 Thin-Plate Splines for Feature Extraction

Thin-plate splines provide a good method to parameterize a warping transformation based on a set of fixed points. It effectively generates a minimal energy solution to a point constrained warping. This lends itself quite nicely to facial affect analysis because the facial affects can be thought of as the deviation of facial action units from a neutral zero-energy position. We show that by selecting landmark points that correspond to separate action unit areas, a good statistic for affect analysis can be generated. The thin-plate spline model is also easily separated into an affine portion that describes rigid head movement and a nonlinear portion that describes the warping induced by facial expressions. The classifier then does not need to train for rigid body head motion, allowing for reduced training sets and simplified classification systems.

The formulation of the thin-plate spline model shows this separation of the affine from the nonlinear. We used the same derivation as Bookstein in his paper on principle warps [10]. This model is initialized from the location of the facial feature points in the neutral position. Using a cost function of

$$Z = -U = -r^2 \cdot \log\left(r^2\right)$$
$$r^2 = x^2 + y^2 \tag{1}$$

it can be shown that the function f that is a solution to

$$I_f = \underset{f}{\arg\min} \iint_{R^2} \left(\left(\frac{\partial^2 f}{\partial x^2} \right)^2 + 2\left(\frac{\partial^2 f}{\partial x \partial y} \right)^2 + \left(\frac{\partial^2 f}{\partial x^2} \right)^2 \right) \tag{2}$$

can be written as

$$f = t_1 + a_x x + a_y y + \sum_{i=1}^{n} w_i U\left(\left\| \begin{bmatrix} x_{m,i} \\ y_{m,i} \end{bmatrix} - \begin{bmatrix} x \\ y \end{bmatrix} \right\| \right) \tag{3}$$

where $x_{m,i}$ and $y_{m,i}$ are the i^{th} x and y coordinated from our model. Others have shown that these warping parameters W, T, and A can be calculated by the following equations [9]

$$\begin{bmatrix} W & T & A \end{bmatrix}^T = L^{-1} Y, \tag{4}$$

where L is defined as follows and Y contains the current positions of the tracked points padded with zeros.

$$L = \begin{bmatrix} K & P \\ P^T & 0 \end{bmatrix}, \tag{5}$$

$$P = \begin{bmatrix} 1 & x_{m,1} & y_{m,1} \\ 1 & x_{m,2} & y_{m,2} \\ \vdots & \vdots & \vdots \\ 1 & x_{m,n} & y_{m,n} \end{bmatrix}, \tag{6}$$

$$K_{i,j} = U\left(\left\| \begin{bmatrix} x_{m,i} \\ y_{m,i} \end{bmatrix} - \begin{bmatrix} x_{m,j} \\ y_{m,j} \end{bmatrix} \right\| \right) \cdot \left(1 - \delta_{i,j}\right) \tag{7}$$

Since P and K are computed from the neutral model, L^{-1} only needs to be computed once when the neutral face is initialized. This allows for the fast calculation of the nonlinear warping parameters W as well as the affine warping parameters A.

It is also important to note that even though the affine warping parameters have been separated from the nonlinear parameters, the nonlinear parameters are still dependent on the affine parameters. This can be corrected easily calculating the inverse of the linear portion of the

affine transform and multiplying it with the nonlinear warping parameters W. This effectively removes the dependence on the affine transformation from the nonlinear parameters. This calculation to remove the affine dependency from W in solution S is shown in (8).

$$S = A^{-T} W \tag{8}$$

Thus thin-plate splines provide us with an efficient model for facial affect characterization by providing a closed form solution to the minimum energy warping separated into affine and nonlinear portions. This method does not require iterative techniques or lengthy operations; simply two matrix multiplications and one 2x2 matrix inverse calculation (the inverse of L is precomputed) is sufficient to generate a result. The figures below show examples of the tracked points undergoing a thin-plate spline warping.

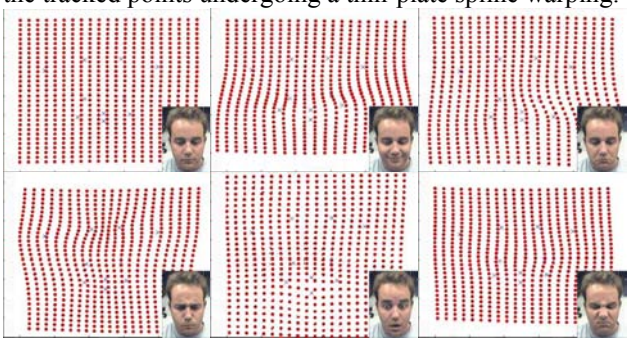

Figure 2 Facial expression feature points (blue Xs) and grid to illustrate warping for neutral, happiness, saddness, anger, surprise, and disgust

Furthermore, a measure of the strength of a particular expression can also be calculated from the thin-plate spline warping parameters. This allows us to not only to distinguish that an expressions is being performed, but also how strong the expression is. The bending norm serves this purpose and is calculated by the equation 9.

$$\mathrm{BN} = \mathrm{trace}\left(\mathrm{W}\,\mathrm{K}\,\mathrm{W}^{\mathrm{T}}\right) \tag{9}$$

It has been shown that affine transformations provide a good approximation to rigid facial movements under planar transformations [11]. By removing this affine component from the feature vector and by selecting facial feature points that are nearly coplanar, we can achieve invariance to rigid body transformations such as head rotations and translations.

3 Real-Time Affect Analysis System (RAAS)

Using the thin-plate spline method of extracting a feature vector, we have developed a real-time system for facial affect analysis. The system is organized into four main components. The initialization routine takes user input to determine the neutral feature templates used in the tracking mechanism as well as the thin-plate spline model parameters. Currently, the initialization is done manually. Other systems have been developed that solve this problem of initialization [16]. Once initialized the program loops through tracking, parameter extraction, and expression recognition. The computational flow is illustrated in Figure 3.

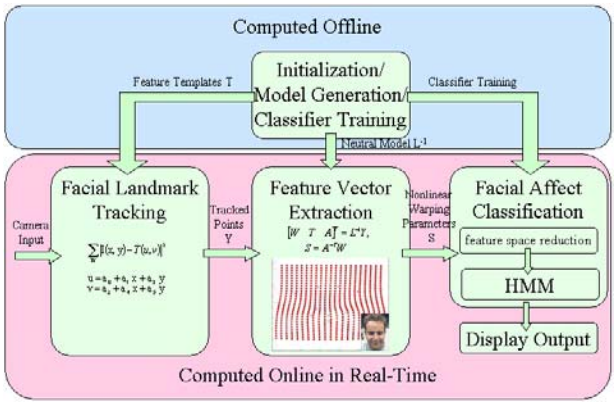

Figure 3: Real-time Affect Analysis System (RAAS) Flow Chart

3.1 Facial Landmark Tracking

In order to track facial landmarks, the system uses a template matching algorithm. This algorithm is similar to that described in [11], but modified for speed. The system uses an affine transformation model described by (10) and (11) to warp the templates in order to track the non-rigid motion.

$$\mathrm{u} = \mathrm{a}_0 + \mathrm{a}_1\,\mathrm{x} + \mathrm{a}_2\,\mathrm{y} \tag{10}$$

$$\mathrm{v} = \mathrm{a}_3 + \mathrm{a}_4\,\mathrm{x} + \mathrm{a}_5\,\mathrm{y} \tag{11}$$

To find a match between the warped template and the current image, we want to find the parameters a_0, a_1, …, a_5 that minimize

$$\sum_{\mathrm{W}} \left|\mathrm{I}(\mathrm{x},\mathrm{y}) - T(u,v)\right|^2 \tag{12}$$

Where u and v are defined as in (9) and (10), I(x,y) is the image intensity at point (x,y), T is the template intensity, and W is the window over which the template is defined.

The system currently tracks ten facial landmarks corresponding to the left and right corners of the lips, the top and bottom of the lips, the left and right nostrils, the outer corners of the eyes and the inner corners of the eyebrows. This provides a sufficiently rich point set from which to collect data while still allowing for the real- time tracking of the feature points.

3.2 Nonlinear Feature Vector Generation

The positions of the facial landmarks are then fed into the thin-plate spline warping parameter calculation. The pre-computed L-1 matrix is multiplied by the current landmark points to yield the affine and nonlinear warping parameters. The nonlinear portion is then modified to remove its dependence on affine transformations and

subsequently used for expression estimation. The details of these calculations are the same as those described in section 3.

In order to provide a better input to the affect estimator portion of the system the warping vector is reduced dimensionally by applying a transformation matrix optimized for linear discrimination via the Ho-Kashyap Algorithm [17]. This generates separating hyper-planes between classes that are then used to project to a lower dimension.

3.3 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 taken from the feature vector produced in the previous step. Using the HMM, a maximum a posteriori estimate of the expression state is computed. [18]

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

$$X = \arg\max_{X_i} P(Y|X_i)P(X_i|X_p) \tag{13}$$

The probability mass function P(X) is taken from the HMM state transition matrix. Specifically P(Xi|Xp) represents the probability of transitioning from state Xp to state Xi. The probability density function (pdf) P(Y|Xi) 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 simple maximum a posteriori estimation provides a fast and efficient means of calculating the current expression.

4 Experimental Evaluation

The thin-plate spline method of feature vector calculation was shown to produce a good metric for various facial expressions. The system was demonstrated to run at about 20 fps on a dual processor system running at 1.5Ghz using a CCD based USB camera running 320x240 resolution.

Two types of experimental studies were conducted in order to validate the performance of the system. In the first tests, the system was tested for its invariance to rotation by training on forward looking images and testing on rotated images. The testing with specific individuals shows the system performance for applications in which the user is known. This demonstrates the systems usefulness in many applications for which user training has been performed. In the second set of tests, the system was trained and tested using the Cohn-Kanade Facial Expression Database [19]. This shows the systems usefulness when applied to individuals for which it has no prior knowledge. Because of the built in measure for strength of expression from the bending norm in the thin-plate spline calculation, we decided that this is a better representation than trying to differentiate between neutral and other expressions for slight movements.

4.1 Invariance to rotation

In order to test the system to rotational invariance, the system was trained on one specific individual using a set of 10 training sets for each expression for a total of 50 training sets. This individual was consistent in making the five facial expressions tested in order to get data that depended more on the relevance of head rotation rather than the different ways people make expressions. The training set was taken with the subject looking straight into the camera with no rotation. Testing samples were then taken with the same subject rotating their head up to 30 degrees in 10-degree increments while performing the trained facial expressions. Ten samples for each expression and orientation were used for testing. None of the testing samples were used as training samples. Head rotations beyond 30-degrees cause tracking errors because tracked points become occluded. Figure 4 shows images demonstrating these tests, followed by confusion matrices demonstrating the results.

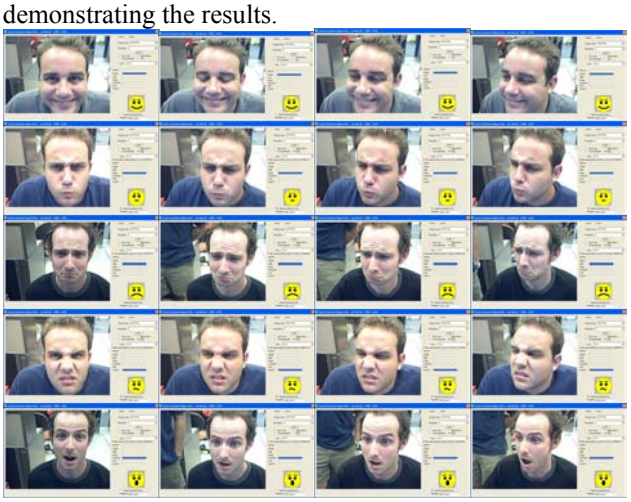

Figure 4 Subjects displaying facial expressions with 0, 10, 20, and 30 degrees of head rotation.

	Happiness	Anger	Sadness	Disgust	Surprise
Happiness	100%	0%	0%	0%	0%
Anger	0%	100%	0%	0%	0%
Sadness	0%	0%	99.7%	0%	0%
Disgust	0%	0%	0%	100%	0%
Surprise	0%	0%	0.3%	0%	100%
No. of Frames in Testing	742	607	673	817	641

Table 1: Test result for 0 degrees of rotation

	Happiness	Anger	Sadness	Disgust	Surprise
Happiness	100%	0%	0%	0%	0%
Anger	0%	99.7%	0%	0%	0%
Sadness	0%	0%	100%	0%	0%
Disgust	0%	0.3%	0%	100%	0%
Surprise	0%	0%	0%	0%	100%
No. of Frames in Testing	872	656	938	877	619

Table 2: Test result for 10 degrees of rotation

	Happiness	Anger	Sadness	Disgust	Surprise
Happiness	100%	0%	0%	0%	0%
Anger	0%	100%	0%	0%	0%
Sadness	0%	0%	99.7%	0%	0%
Disgust	0%	0%	0%	100%	0%
Surprise	0%	0%	0.3%	0%	100%
No. of Frames in Testing	727	704	634	631	513

Table 3: Test result for 20 degrees of rotation

	Happiness	Anger	Sadness	Disgust	Surprise
Happiness	100%	0%	0%	0%	0%
Anger	0%	100%	0.4%	24.4%	0.2%
Sadness	0%	0%	93.8%	0.6%	0%
Disgust	0%	0%	0.3%	75.6%	3.2%
Surprise	0%	0%	5.9%	0%	96.6%
No. of Frames in Testing	636	493	597	640	530

Table 4: Test result for 30 degrees of rotation

We can see from the results that the feature vector does well in removing the rigid body motion until about 30 degrees. In this situation the fact that the face points are not actually on the same plane causes some extra nonlinear warping which deteriorates the accuracy of the system. The extremely high accuracy of the system in this case can be attributed to the training and testing on the same person as well as the consistency in which that person performed the facial expressions. In a real-world environment, which is difficult to reproduce in a laboratory setting, these results might be different. The training data was chosen to include some amount of noise in the landmark tracking in order to be able to correctly classify when faced with noisy tracking.

4.2 Generalization

The Cohn-Kanade Facial Expression Database was used to test how well the system generalizes to a variety of individuals without having to retrain the system for each person. First a subset of the data was taken for which the corresponding FACS codes communicate one of the 5 facial expressions classified by the system. Of these samples, a training set of 20 samples for each expression was taken which was not used for testing. The classification was iterated 1000 times with randomly chosen training sets and the results were averaged. In total, 295 samples of facial expressions taken from 95 different subjects were used to evaluate the system (74 happiness, 39 anger, 40 disgust, 71 sadness, and 71 surprise). In order to test the short video samples, the last three frames were classified and compared to the ground truth data. The last three frames were chosen because some of the sequences are only a few frames long and the last few frames generally correspond to the highest intensity of expression. The samples contained variations in lighting, head movements, and ethnicity [19]. It should also be pointed out that the actors in the database were asked to mimic expressions shown to them by an experimenter. The expressions in this database were also therefore somewhat exaggerated from real-life expressions and might not be well representative of real-life behavior, but the data is sufficient to show the viability of the system to detect changes in facial expressions.

It has been pointed out that because of the large variations in the action units used by various individuals to express the six universal facial expressions it is better to perform analysis on the FACS codes themselves [8]. We chose to implement the system to detect expressions based on the same core FACS codes. The system could also be trained to recognize individual FACS codes. Testing was performed on those expressions for which there were a sufficient sample size of expressions with correspondingly similar FACS codes. Because of the wide range of FACS codes that are used to express fear, this expression was not included in the testing.

Table 5 shows the confusion matrix for the different facial expressions from the Cohn-Kanade Facial Expression Database. The columns represent the ground truth and the rows show the classification results.

	Happiness	Anger	Sadness	Disgust	Surprise
Happiness	95.91%	0.86%	0.61%	0.15%	1.30%
Anger	1.60%	70.09%	13.53%	15.13%	0.59%
Sadness	0.14%	11.93%	73.09%	6.23%	1.09%
Disgust	0.17%	16.13%	9.75%	72.15%	0.59%
Surprise	2.18%	0.58%	3.02%	6.33%	96.43%
No. of Frames in Testing (# of subjects)	117 (39)	12 (4)	108 (36)	15 (5)	108 (36)

Table 5: Person-independence Test Results

We can see from these results that the system works best on expressions that involve large movements. The expressions of anger, disgust and sadness perform worse

because of the small movements of the tracked points combined with the similarity of the expressions themselves. This can also be seen in the similarity of the thin-plate spline warpings shown in Figure 2; this is especially true for anger and disgust.

5 Concluding Remarks

Facial expression recognition is complicated by a large number of factors. These factors can include variations in illumination, changes in head position, and variations in the way different people express emotions. In order to compensate for these obstacles, we took the approach of attempting to extract a feature vector that is invariant to head movements, but still contained the information necessary to accurately classify facial expressions. We accomplished this by using thin-plate splines to extract a feature vector and applied this technique in a real-time affect analysis system. The results show the effectiveness of the thin-plate spline feature vector as a rotation invariant measure of nonlinear facial movements. Furthermore, it was shown that this technique can be easily applied in a fast and efficient manner sufficient for real-time system operation. Testing of the system for independence between persons has also shown that it is capable of working well even without having been trained on a specific individual.

References

[1] K. Huang, M. M. Trivedi, Driver Head Pose and View Estimation with Single Omnidirectional Video Stream, Proceedings of the 1st International Workshop on In-Vehicle Cognitive Computer vision Systems, in conjunction with the 3rd International Conference on Computer Vision Systems, Graz, Austria, April 3, 2003.

[2] B. Fasel, J. Luettin, "Automatic facial expression analysis: a survey", Pattern Recognition, Vol. 36, pages 259-275, 2003.

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

[4] J.N. Bassili, 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. 1979.

[5] 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.

[6] I. A. Essa, and A. Pentland, Facial Expression Recognition using Image Motion. Motion Based Recognition , M. Shah and R. Jain (Editors), Kluwer Academic Publishers, Computational Imaging and Vision Series, 1997.

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

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

[9] J. Gower, Generalized Procrustes Analysis. Psychometrika, 40:33--51, 1975.

[10] F. L. Bookstein, Principle Warps: Thin-Plate Splines and the decomposition of deformations, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 11, No. 6, June 1989.

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

[12] H. Hong, H. Neven, and C. von der Malsburg, Online Facial Expression Recognition based on Personal Galleries. Intl. Conference on Automatic Face and Gesture Recognition, 1998.

[13] C. Lisetti, and D. Rumelhart, Facial Expression Recognition using a Neural Network, 11th International Flairs Conference, AAAI Press, 1998.

[14] M. Dumas, Emotional Expression Recognition using Support Vector Machines, Technical Report, UCSD, MPL, 2001.

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

[16] L. Wiskott, J.-M. Fellous, N. Kruger, and C. von der Malsburg, Face recognition by elastic bunch graph matching, Tech. Rep. IR-INI 96-08, Institut fur Neuroinformatik, Ruhr-Universitat Bochum, D44780 Bochum, Germany, 1996.

[17] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification (Second ed.) John Wiley & Sons, Inc., New York, 2000.

[18] L. R. Rabiner, A tutorial on Hidden Markov Models and selected applications in speech recognition. Proceedings of the IEEE, 27(2):257-286, 1989.

[19] T. Kanade, J. F. Cohn, and Y. Tian, Comprehensive Database for Facial Expression Analysis. Proceedings of the 4th IEEE International Conference on Automatic Face and Gesture Recognition (FG'00), March, pp. 46 – 53, 2000.

[20] J. McCall, S. Mallick, M. M. Trivedi, Real-Time Driver Affect Analysis and Tele-viewing System, Intelligent Vehicles Symposium, Proceedings. IEEE , June 9-11, 372 - 377, 2003.

[21] I. Cohen, N. Sebe, A. Garg, L. S. Chen and T. S., Huang "Facial expression recognition from video sequences: temporal and static modeling", Computer Vision and Image Understanding, Vol. 91, Pages 160-187, 2003.