
Image-based stress recognition from a model-based face tracking system

Sundara Venkataraman

SUNDARA@PAUL.RUTGERS.EDU

CBIM, Department of Computer and Information Sciences, Rutgers University, Piscataway, NJ 08854 USA

Abstract

This paper presents a comparison of learning methods to detect stress from video sequences of people with a close-up of their face. The video sequences consist of people subjected to various psychological tests that induce high and low stress situations. We use a model-based tracking system to get the face movements and deformations in a parameterized form. We compare different learning methods to learn from these parameters to do recognition of high and low stress situations for labeled video sequences. We will present results of using Hidden Markov Models (HMMs) for recognition. The main contribution of this paper is a novel method of stress detection from image sequences of a person's face.

1. Introduction

Stress detection of humans has been a well researched topic in the area of speech signal processing [Steeneken, Hansen 99], while very little attention has been paid to recognizing stress from faces. Recognizing stress from faces could complement speech-based techniques and also help in understanding recognition of emotions. In the next two sections we illustrate the data collection procedure and the algorithms that will be used for training/recognition.

1.1 Overview of the system

The data used for our experiments were obtained from a psychological study at the University of Pennsylvania. The subjects of the study were put through a battery of tests that induce high and low stress situations. The subjects were videotaped as they took the tests.

A generic model of the face is fitted to the subjects' face and the tracking system is run on the face image sequence with this initial fit of the model in the first frame. The tracking system does a statistical cue integration from computer vision primitives such as edges, point trackers and optical flow. The face model incorporates some parametric deformations that give jaw, eyebrow and basic lip movements. The face tracker gives values for these parametric deformations as a result of the tracking. These

are the parameters we will use to learn the movements that correspond to different stress situations. The learning methods will be trained on these parameters and the parameters of a sequence from an unknown stress situation are tested against the learned system to classify a given sequence as a high or low stress condition.

1.2 Learning and recognition

We will evaluate different learning approaches to train the system for low and high stress conditions. The tracking result will give us parameters which account for the rigid transformations between the face movements and also deformations of the mouth, eyebrow etc. Hidden Markov Models (HMMs) are used to recognize stress conditions of a video sequence as a high or low stress situation. HMMs were chosen since they have been used to model temporal dependence very effectively in American Sign Language (ASL) recognition.

2. Deformable model tracking

2.1 Deformable models

The face tracking system uses a face model that deforms according to movement of a given subject's face. So the shape, position and orientation of the model surface can change. These changes are controlled by a set of n parameters \mathbf{q} . For every point i on the model surface, there is a function F_i that takes the deformation parameters and finds

$$p_i = F_i(\mathbf{q})$$

where p_i is the position of the point in the world frame [Metaxas 97].

Most computer vision applications such as deformable model tracking, require the first order derivatives, so we restrict F_i to the class of functions for which the first order derivative exists everywhere with respect to the parameter \mathbf{q} . This derivative then is the Jacobian J_i , where

$$J_i = \begin{bmatrix} | & & | \\ \frac{\partial p_i}{\partial q_1} & \dots & \frac{\partial p_i}{\partial q_n} \\ | & & | \end{bmatrix}$$

Every column l of the Jacobian matrix \mathbf{J}_i is the gradient of p_i with respect to the parameter q_l .

2.2 Fitting and Tracking

In principle, there exists a clean and straightforward approach to track deformable model parameters across image sequences. Low-level computer vision algorithms generate desired 2D displacements on selected points on the model (differences between where the points are currently according to the deformable model and where they should be according to measurements from the image. These displacements, also called ‘image forces’, are then converted to a single n -dimensional displacement \mathbf{f}_g in the parameter space, called the *generalized force* and used as a force in the first-order massless Lagrangian system :

$$\dot{\mathbf{q}} = \mathbf{f}_g + F_{\text{internal}}(\mathbf{q}) \quad (1)$$

where $F_{\text{internal}}(\mathbf{q})$ is the result of internal forces of the model (i.e. elasticity of the model, preset). This system is integrated using the classical Euler integration procedure, which eventually yields a fixed point, where $\mathbf{f}_g = \mathbf{0}$. The fixed point thus obtained corresponds to the desired new position of the model.

In order to use the system of (1), all the 2D image forces from the computer vision algorithms have to be accumulated into \mathbf{f}_g . First, we convert each image force \mathbf{f}_i on a point p_i into a generalized force \mathbf{f}_{gi} in parameter space, which describes the effect that a single displacement at the point p_i has on all the parameters. Then, obtaining a generalized force \mathbf{f}_g simply amounts to summing up all \mathbf{f}_{gi} .

$$\mathbf{f}_g = \sum_i \mathbf{f}_{gi} \quad \text{where} \quad \mathbf{f}_{gi} = \sum_i \mathbf{B}_i^T \mathbf{f}_i$$

and

$$\mathbf{B}_i = \left. \frac{\partial \mathbf{Proj}}{\partial p} \right|_{p_i} \mathbf{J}_i \quad (2)$$

\mathbf{B}_i is the projection of the Jacobian \mathbf{J}_i from world coordinates to image coordinates through the projection matrix \mathbf{Proj} at point p_i .

Generating the generalized force this way works fine as long as all the image forces come from the same cue (diff. algorithms on the same image). When there are multiple cues from multiple vision algorithms, combining the cues becomes a hard problem. In order to effectively combine them statistically we will need to know the distributions of the individual cues ahead of time, but it is hard to estimate these distributions beforehand.

The affine regions framework [Goldenstein et al. 2001, 2003] is chosen to estimate the distributions of the cues within small regions and the equivalent of the central limit theorem is applied to these affine regions to make a Gaussian approximation. Then we use maximum likelihood estimation to get the final generalized force.

The face model itself was made from a publicly available geometric model of the head, available from the University of Washington as part of [Pighin et al.99]. A face mask was cut out of this original model and we obtained a static model with 1,100 nodes and 2000 faces. Then, parameters and associated regions are defined for the raising and lowering of eyebrows, for the smiling and stretching of the mouth, for the opening of the jaw as well as the rigid transformation parameters for the model frame.

In this version of the system we have added asymmetric deformations for the eyebrows and the mouth region i.e. the left and right eyebrows, the left and right ends of the lips of the mouth are no longer tied together. This is essential since one of the major indicators of stress is asymmetric lip movements and the original framework did not support that.

The deformation parameters all put together form about 14 parameters. The tracking results from a particular video sequence will give us these 14 parameters for the model for each time instance. We perform these tracking experiments on various subjects and use the parameters we obtain to train the HMMs.

2. Hidden Markov Models

3.1 Background

Hidden Markov Models (HMMs) have been a very effective tool in capturing temporal dependencies in data and fitting them to models. They have been applied very effectively to the problem of American Sign Language (ASL) recognition and to speech recognition in particular with reasonable commercial success.

Detecting stress patterns is very similar to other tasks where recognizing activities is the main goal such as gesture recognition, sign language recognition etc. For such tasks, HMMs have been shown to be highly suitable, for several reasons : First, the Viterbi decoding segments the data into its components implicitly, which solves the problem of data segmentation. Second, the state-based nature of HMMs is an ideal system for recognizing signals over a period of time. Third, the statistical nature of HMMs makes them ideal for recognizing tasks which have some inherent variation in them.

A more detailed introduction and description of algorithms and inference in HMMs is described in [Rabiner 89].

4. Experiments and results

The data used in the experiment were deformation parameters of the face model from the tracking system. The asymmetries in these deformation parameters are a prominent indicator of stress (ex. eyebrow movements, curving lips etc.). The learning system is trained with these asymmetries to recognize high stress situations.

Figures 1-6 show the various parameters used in the learning process.

HTK is a toolkit for creating and manipulating HMMs. HTK was chosen over BNT due to ease of use for our present application. The deformation parameters obtained over time are fed into the HMM system created by configuring the HTK.

The training system consists of two HMMs, one each for high and low stress situations. Each HMM consists of 5 states.

A number of subjects were subjected to high and low stress situations and were videotaped during this process. This labeled data was split into training and testing data for the experiment.

A 75%-25% split between training and testing data gave 100% recognition with all the 6 samples of the test set being identified correctly as low/high stress situations. A 50%-50% split between training and testing data gave 92% recognition with 12 of the 13 test samples being recognized correctly.

3. Conclusions and future work

The structure and probabilities of a Bayesian network to do recognition needs to be investigated further. If the recognition rate does poorly, boosting could be used to

combine these two classifiers and would make the recognition more robust.

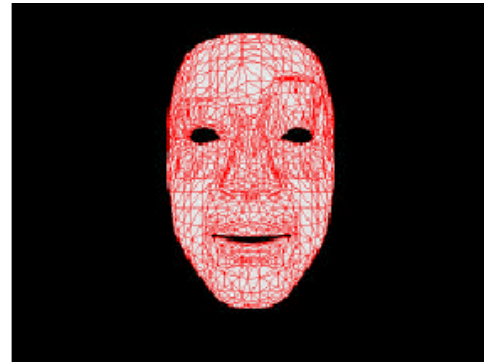


Figure 1. Left eyebrow movement

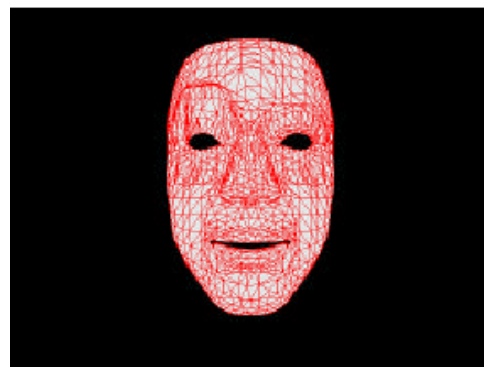


Figure 2. Right eyebrow movement

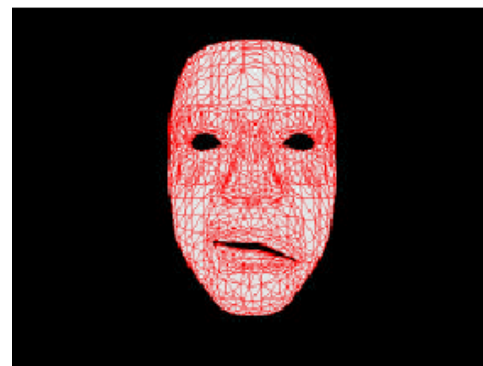


Figure 3. Left risorius movement

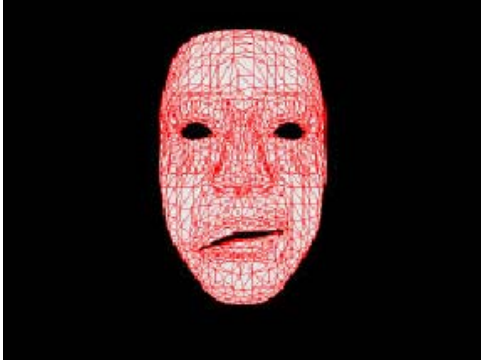


Figure 4. Right risorius movement

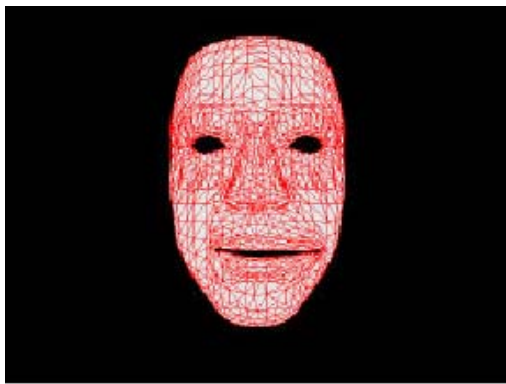


Figure 5. Left lip stretching

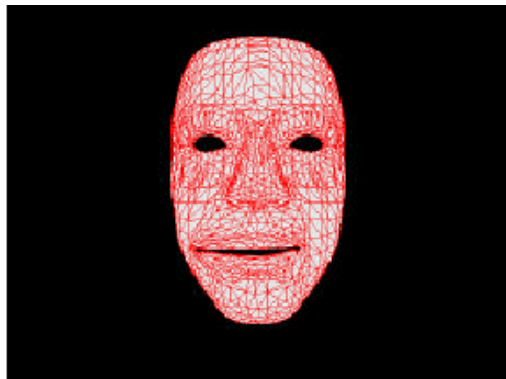


Figure 6. Right lip stretching

References

Steeneken, H. J. M., Hansen, J. H. L. (1999). *Speech under stress conditions: Overview of the effect on speech production and on system performance*. IEEE

International Conference on Acoustics, Speech and Signal Processing.

Goldenstein, S., Vogler C., Metaxas D.N., (2001). *Affine arithmetic based estimation of cue distributions in deformable model*.

Goldenstein, S., Vogler C., Metaxas D.N., (2003). *Statistical Cue Integration in DAG Deformable Models*. IEEE Transactions on Pattern Analysis and Machine Intelligence, July 2003.

Vogler, C., Metaxas, D. N. (1999). *Parallel Hidden Markov Models for American Sign Language recognition*. International Conference on Computer vision, Kerkyra, Greece.

Pighin, F., Szeliski, R., Salesin, D., (1999). *Resynthesizing Facial animation through 3D Model-based tracking*. International Conference on computer vision.

Rabiner, L. R., (1989). *A tutorial on hidden Markov models and selected applications in speech recognition*. Proceedings of the IEEE, Vol. 77.

Metaxas, D. N., (1997). *Physics-based deformable models: Applications to computer vision, graphics and medical imaging*. Kluwer Academic Press.

The Hidden Markov Model toolkit : <http://htk.eng.cam.ac.uk/>