Facial Action Coding Using Multiple Visual Cues and a Hierarchy of Particle Filters

Joel C. McCall and Mohan M. Trivedi
Computer Vision and Robotics Research Laboratory
University of California, San Diego
jmccall@ucsd.edu mtrivedi@ucsd.edu

Abstract

In this paper we present a framework for tracking non-rigid facial landmarks by combining various visual cues at multiple levels of detail. Using a probabilistic framework consisting of a hierarchy of particle filters, we are able to track individual facial landmarks using multiple visual cues at the local level, as well as tracking results at more coarse level of detail. This allows for the fusion of global and local cues in an efficient and robust manner. Testing is performed by tracking and classifying facial action codes obtained from the Cohn-Kanade AU-Coded Facial Expression Database.

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 [16]. It has been shown that there are six universally common facial expressions for displaying the emotions of Anger, Disgust, Fear, Happiness, Sadness, and Surprise [7]. However, A more descriptive alphabet of human emotion can be constructed by looking at individual components of facial expressions; specifically, we implore facial action units as presented by Ekman et al [8] as a method for identify emotion and mental state. We will help to identify these facial action units by examining the motion of individual facial landmarks such as corners of the eyes, mouth, eyebrows, etc. Classifying individual action units allows us to explore other areas of human computer interactions by broadening the number of emotional states that can be captured. This thereby expands the contexts in which such systems can be used. As an example, intelligent vehicle systems for driver departure warning can be enhanced with the ability to predict fatigue and attentiveness [16].

Facial expression and affect analysis is something we instinctively perceive; however, many factors contribute to the difficulty in machine identification of facial actions and expressions. These factors include difficulties imposed by lighting conditions, varying emotional factors that lead to facial expressions, variations in the expressions between persons, head poses and head movements [11].

In this paper we propose a novel framework that integrates the problem of face and facial landmark detection and tracking using a variety of cues. By constructing a probabilistic model based on a hierarchical pyramid of facial region trackers, we can more robustly predict individual facial landmark locations while at the same time provide robustness to occlusion. In our pyramid of facial region trackers, we first detect and track faces in a given image. Using this observed knowledge of the location of faces within the image, we can more easily and accurately detect smaller regions of the face. This continues until we have identified specific facial landmarks. We will discuss this approach in more detail in section 2. The tracking mechanisms used in each layer of the hierarchy are constructed using particle filtering [15]. Particle filtering provides a method for tracking complex distributions over time in an efficient manner as well as combining observations from a variety of cues. By fusing multiple cues we can create robustness to a variety of environmental and lighting conditions. For example, changes in lighting might cause cues such as facial motion to yield poor results as algorithm assumptions are violated, but cues retrieved from facial structure are often more robust to such conditions.

1.1. Related Research in Facial Affect Analysis

In order to robustly identify facial actions, one must first identify the subjects face as well as properly register the location of specific facial landmarks. From this point on a variety of techniques can be employed to analyze facial expressions and facial action codes.

Bassili [1] has shown the humans are capable of identify-

Proceedings of the 2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06)
0-7695-2646-2/06 $20.00 © 2006 IEEE

ing expressions when presented simply with points placed at facial landmarks. Optical flow and facial motion has also been shown to be useful in the automatic identification of facial affects and expressions [9, 10, 18]. However, these methods often break down when rigid body motion, due to head movement, is present. Attempts to solve this problem of separating the rigid motion from the non-rigid motion have been made using model-based estimation [9].

Non-rigid feature tracking has been explored in a variety of ways ranging from methods based purely on optical flow and probabilities, to methods involving the construction detailed facial models and perturbing them according to facial landmark movement. Black and Yacoob parameterized various facial feature motions with affine and similar transformation models [2]. Another approach to solve this problem is to input more complex feature vectors into classification systems. Systems developed using Graph Matching [14], Neural Networks [20], and Support Vector Machines [6] have been shown to be effective, but require more complex classification schemes.

Appearance cues have also been shown to be very important for facial affect identification. Active appearance models [21] use a linear combination of detailed 3-dimensional models to estimate head pose and appearance. The fitting is performed by aligning a model generated image with the input image. Generic appearance-based object detectors based on Haar wavelets and boosted classifiers have also been shown to be useful in detecting faces [29] and facial features [12]. These types of single frame detectors have also been included as observations to tracking systems [23].

Particle filtering provides a robust method for tracking objects and features by estimating the probability of the object's current state given previous states and observations [15]. It accomplishes this using Monte Carlo sampling techniques to approximate probability density functions without any assumptions of Gaussianity. Others have used combinations of particle filters for tracking facial features and constraining them to fit particular models of facial movement [27].

In our work, we build upon this previous research by developing a hierarchical structure of particle filters which operate on a variety of different observations. This hierarchy allows for the coarse to fine tracking of facial landmarks. Region tracking results obtained from coarser levels of detail are used to condition the probability estimates of the finer detailed face regions. The observations at each level of the particle filter hierarchy include structural information, relative orientations, locations, and sizes to other filters in the hierarchy; general appearance information, generated by Haar wavelet based object detection; and specific appearance information, generated by adaptive templates of facial regions and landmarks.

2. Real-time Affect Analysis using Hierarchical Particle Filtering

In this section we will introduce the components that make up our Real-Time Affect Analysis System (RAAS). This system is composed of multiple levels of particle filters related through a hierarchical structure that allows the propagation of landmark location estimates from a coarse resolution to a fine resolution. Figure 2 consists of a diagram showing the structure of this system. The individual components will be described in the next two sections of this paper.

2.1. Particle Filtering Overview

1) If Resampling, generate new samples $\mathbf{s}_t$ from the samples $\mathbf{s}_{t-1}$ using a multinomial distribution with coefficients corresponding to the weights $\mathbf{w}_{t-1}$. Reset weights $w_{i,t-1} = 1/N \forall i \in \{1, ..., N\}$.

2) Update each sample i using the transition prior such that
$s'_{t,i} = p\left(x_t | x_{t-1} = s_{t,i}\right)$

3) Re-weight the samples based on the current observations using the equation
$w_{t,i} = w_{t-1,i} p\left(y_t | x_{t,i} = s'_{t,i}\right)$

Figure 1. Particle Filtering Algorithm

Particle filtering is a convenient method for estimating the probability density of the current state of an object given the object's previous states as well as current and past observations. this is expressed mathematically as $p\left(x_t | x_{0:t-1}, y_{0:t}\right)$, where x_t is the object state at time t and $y_{0:t}$ are the observations from times 0 through t. This distribution can be estimated using a properly weighted sample distribution. By choosing a transition prior of $p\left(x_t | x_{t-1}\right)$, the weight update procedure for each sample, indexed by i at time t, simply becomes $w_{t,i} = w_{t-1,i} \cdot p\left(y_t | x_{t,i}\right)$ [25]. The filtering algorithm used in each of the levels of our system is the same as that presented in [15] and shown in figure 1.

2.2. Creating a Hierarchy of Particle Filters

One limitation of particle filtering is its susceptibility to the "curse of dimentionality," [5] where the number of samples required for an accurate estimation grows exponentially with the dimension of the state space. For facial features, the dimensionality for the complete characterization of facial motions is quite large, requiring the state space to be split into multiple filters which can be constrained based on a facial model. Others have done this by separating facial

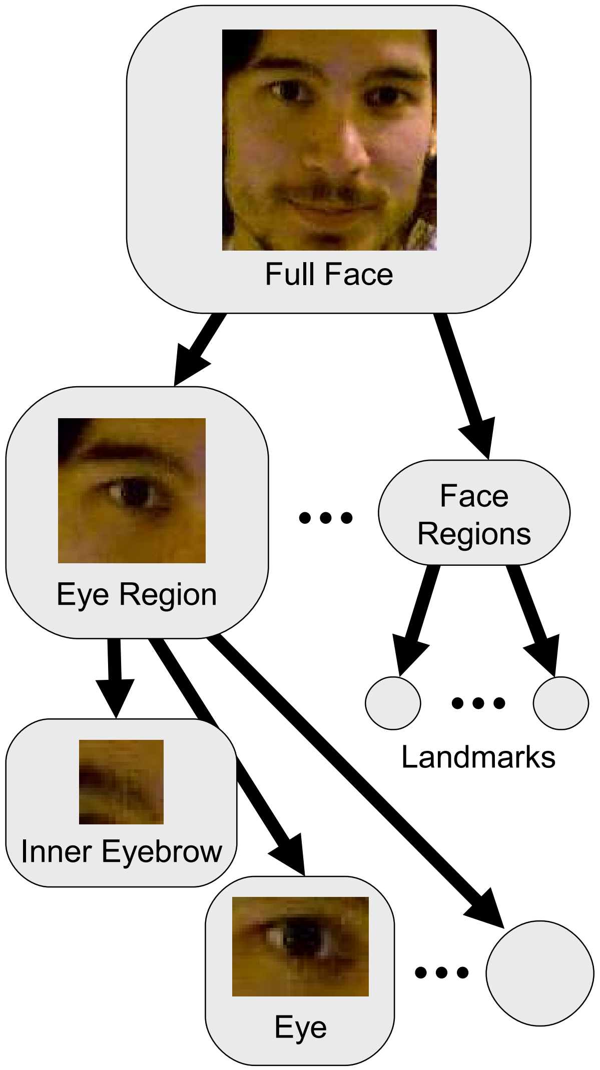

Figure 2. Bayesian network for facial landmark tracking. The structure for the right eye landmark tracking is emphasized. The associated probabilities are propagated in time using particle filtering to make the network dynamic.

landmarks into groups and then tracking the groups individually under a constrained facial model [27].

By using a dynamic Bayesian network framework, we can condition the finer detailed tracking on results achieved from tracking regions at a coarse level. The structure of our Bayesian network is shown in figure 2. Each node in the network is made dynamic by propagating the probabilities in time using particle filtering. Each node also has it's own observations as we will describe later. As an example, the subset of the network for location of eye region landmarks highlighted in the figure is expressed in mathematical terms in Equation 1. Where EB, E, ER, F are the parameterizations for the eyebrow, eye, eye region, and face respectively.

$$p\left(EB, E, ER, F\right) = p\left(EB|ER\right)p\left(E|ER\right)p\left(ER|F\right) \tag{1}$$

Using this framework, we can adjust the particle filtering mechanism in each of the nodes by conditioning them on their respective parent nodes. Our particle filter now tries to find the probability $p\left(x_t|x_{0:t-1}, y_{0:t}, z_t\right)$ where z_t represents the parent node's parameters. This effectively adds an additional observation to the nodes particle filter. Equation 2 expresses the sample re-weighting step (step 3 in figure 1) using this modified filter.

$$w_{t,i} = w_{t-1,i} \cdot p\left(z_t, y_t|x_{t-1,i}\right) \tag{2}$$

The parameters we chose to represent each node consists of the x and y image coordinates of the region, the size of the region, the aspect ratio of the region, and the rotation of the region along the image axis. Our state transition probability was chosen to model additive noise with zero mean. The variance was set to account for non-linear facial movements and was determined through empirical testing. More complex models could be trained from facial motion data. In the next section we will explain the node observations in more detail.

3. Fusion of Multiple Observations

Particle filtering provides a useful probabilistic framework for incorporating multiple observations. Using the "naive" Bayesian assumption that the observations are independent of each other, the observation probability density function can be factored into the product of each of the density functions of the observations o contained in the set of all observations Obs. Strictly, the naive Bayesian assumption doesn't hold; however, in practice systems based on this assumption have been shown to work quite well [4]. This density function is shown in equation 3.

$$p\left(y_t|x_t\right) = \prod_{o \in Obs} p_o\left(y_{t,o}|x_t\right) \tag{3}$$

The manner in which we choose these probability density functions can provide flexibility and robustness to the system. For example, certain observations, such as those generated from adaptive templates, initially might not provide good information. In this case, we can assign a high variance Gaussian distribution or even a uniform distribution to the density function so that it has little or no impact on the sample re-weighting. We can perform similar adjustments in cases where certain features are occluded, relying on structural information and observational outputs of the other filters in the current and higher levels in our particle filter hierarchy.

The observations at each level of the particle filter hierarchy include general appearance information, generated by Haar wavelet based object detection; person-specific appearance information, generated by adaptive templates of facial regions and landmarks; and global structure cues; generated by the higher order filter of the hierarchy. In the sections below we will describe these cues currently used in the RAAS system.

3.1. Haar Wavelet Based Cues

Object detection using a cascade of boosted Haar wavelet based classifiers have been shown to provide highly accurate results with little false positives [29]. Others have used the output of Haar wavelet classifiers in filtering systems [23]. Similarly, in our system, we construct the observation probability by constructing a mixture of equally weighted Gaussian distributions centered around each detected object. The cascade of classifiers are trained on faces sampled from the FERET database [24] which we hand labeled with the locations of individual facial landmarks. The cascades for the top level particle filter was trained against randomly selected non-face background images obtained from the internet. The cascades for lower level particle filters were trained against background images of facial regions excluding the region or landmark being detected.

To deal with occlusion of regions of the face, multiple Haar wavelet observers can be combined a la the "naive" Bayesian assumption stated earlier at each level of our hierarchy. For example, the top level face detection filter uses Haar wavelet detectors for the eye region, nose region, and mouth region separately, allowing for partial occlusion of the face and providing more robustness to false positives generated by any single detector. This approach of using a naive Bayesian classifier on different facial regions is similar to that developed by Schneiderman and Kanade [26]. This is expressed mathematically in equation 4.

$$\begin{aligned} p\left(y_{t,face}|x_t\right) \quad = \quad & p\left(y_{t,eyes}|x_t\right) \\ & p\left(y_{t,nose}|x_t\right) \\ & p\left(y_{t,mouth}|x_t\right) \\ & \prod_{o\in Obs} p_o\left(y_{t,o}|x_t\right) \end{aligned} \tag{4}$$

3.2. Adaptive Template Cues

Person-specific appearance cues can also be used rather than relying solely on appearance models generated from facial database which might not include the current observed subject. In our system, we construct adaptive templates based on a IIR filtering of the detected facial regions and landmarks. These appearance models are initialized using the average appearance of the corresponding facial regions and landmarks seen in the FERET database used in constructing the general appearance cues. After each frame, the template is updated by taking a weighted sum of the current region or landmark appearance and the adaptive template itself. The weighting represents the responsiveness of the template to change.

The observation density function for the adaptive template is assumed to be a zero-mean Gaussian in the sum-of-squared pixel error between the sample and the template. By measuring the minimum of the sum-of-squared error over all of the samples, we can get a measure of the performance of the adaptive template. This is useful in determining the variance of the observation probability. Higher variances will place less emphasis on the associated observations. This is because the higher the variance, the greater the entropy of the observation function and the closer the density gets to uniform over the samples. Observations having a uniform distribution over all of the samples have no impact on the weight updates as all of the weights are updated by the same value. Figure 3 shows the initial templates for a few of the face regions and landmarks.

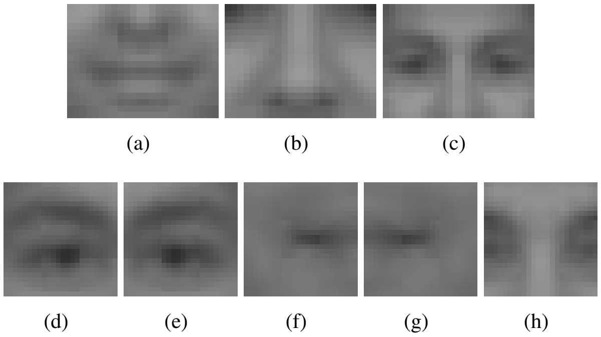

Figure 3. Face Region Templates: (a) Mouth Region, (b) Nose Region, (c) Eye Region, (d) Right Eye, (e) Left Eye, (f) Right Mouth Corner, (g) Left Mouth Corner, (h) Nose Bridge

3.3. Higher Level Cues

As demonstrated in equation 2, we can view the information from the higher level regions as observation to the lower regions. Assuming conditional independence between the various node observations and the higher level observations, we can write our node observation density function (step 3 in figure 1) as

$$\begin{aligned} p\left(z_t, y_t|x_t\right) \quad = \quad & p\left(z_t|x_t\right) \\ & p\left(y_{t,haar}|x_t\right) \\ & p\left(y_{t,template}|x_t\right) \end{aligned} \tag{5}$$

where z_t represents the parameters of the higher level region. The density function $p\left(z_t|x_t\right)$ is assumed gaussian and learned from training data associated with the relative location of the higher level region to the currently tracked region.

4. Incorporation with Facial Action Code detection

Thin-plate splines have been shown to provide a good feature vector for facial expression classification [19, 22]. This method also works well with our system in that the tracked landmarks can be used for control points in the thin-plate spline warping. Thin-plate splines furthermore have the usefulness of parameterizing the warping into affine and non-linear portions which is useful for creating robustness to rotation and translation. In our system, we generate a feature vector based on this non-linear warping and input this feature into an AdaBoost classifier [13] using a decision stump as the weak learner.

5. Results

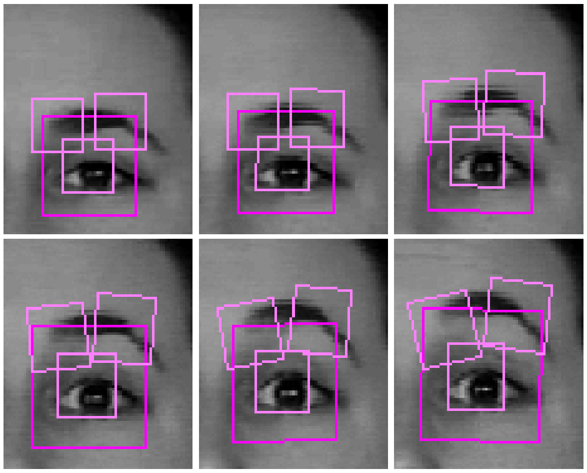

Figure 4. Tracking results for the eye region and subregions. The lighter boxes denote subregions of the eye region.

Testing of the system was performed using the Cohn-Kanade AU-Coded Facial Expression Database [17]. The database consists of 97 subjects each performing a variety of coached facial expressions. Because of the difficulty in generating natural facial expression in a laboratory setting, it is easier to test system performance using Facial Action Codes (FACS) [28]. As described in section 1, FACS can also be considered the building blocks for facial expressions as they can be attributed to specific muscles or muscle groups within the face. The database contains ground truth information about the specific FACS which are present in each of the sequences. Figure 4 shows the results of the tracking on one of the test subject performing the surprised expression.

Of the 97 subjects in the database, the system successfully initialized on all but one subject. Table 1 shows the results for some of the FACS associated with the upper portion of the face. The results were obtained by training on a randomly generated set of sequences and testing on the remaining set of sequences. This procedure was repeated 100 times and the results were averaged. A sequence is considered a positive sequence if the specific action unit is expressed in any part of the sequence, a negative sequence otherwise. The metrics shown in the table include the overall accuracy (percentage of correct test sequences), the detection rate (percentage of positive sequences detected), and the false alarm rate (percentage of negative sequences incorrectly identified). All sequences (both training and testing) were tracked using the method described in this paper to generate the feature vectors for classification.

6. Conclusions

In this paper we have shown a novel framework for detecting and tracking faces and facial features. Using a hierarchy of filters, we can create robustness to noise and occlusion. Facial landmark tracking is improved by using the prior information of more coarse levels of details, such as the location of the face or specific facial regions. Finally we demonstrated the value of this framework by applying it to facial action code recognition using a standard database for facial expressions. This type of framework has far reaching interest in articulated body tracking and human-computer interface applications. [3]

Acknowledgment

Research reported in this paper was sponsored by a UC Discovery Grant and Digital Media Innovations (DiMI). Special thanks to our colleagues at the Computer Vision and Robotics Research laboratories, Dr. Douglas Fidaleo, and Professors Cohn and Kanade.

References

[1] 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.

[2] 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.

[3] J.-Y. Bouguet. Camera calibration toolbox for matlab.

[4] P. Domingos and M. Pazzani. Beyond independence: Conditions for the optimality of the simple bayesian classifier. In *Proceedings of the Thirteenth International Conference on Machine Learning*, pages 105–112, 1996.

[5] R. Duda, P. Hart, and D. Stork. *Pattern Classification (Second ed.)*. John Wiley & Sons, Inc., New York, 2000.

Table 1. Test Results for Facial Action Code Classification

	Facial Action Code					
	1	2	4	5	6	7
Description	inner brow raised	outer brow raised	brows lowered	upper eyelids raised	cheeks raised	lower eyelids raised
Overall Accuracy	86.3%	88.9%	80.2%	84.0%	80.9%	79.6%
Detection Rate	88.2%	85.3%	76.0%	76.6%	76.0%	71.4%
False Alarm Rate	15.7%	7.5%	15.6%	8.6%	14.3%	12.2%
Training Sequences (# positive)	457 (132)	457 (85)	457 (140)	457 (67)	457 (100)	457 (93)
Testing Sequences (# positive)	20 (10)	20 (10)	20 (10)	20 (10)	20 (10)	20 (10)

[6] M. Dumas. Emotional expression recognition using support vector machines. Technical report, Machine Perception Lab, Univeristy of California, San Diego, 2001.

[7] P. Ekman. Expressions of emotion: An old controversy and new findings. *Philosophical Transactions of the Royal Society of London, Series B*, 335(1273):63–69, 1992.

[8] P. Ekman and W. V. Freisen. The facial action coding system: A technique for measurement of facial movement, 1978.

[9] I. A. Essa and A. Pentland. Facial expression recognition using a dynamic model and motion energy. In *Proceedings of the International Conference on Computer Vision*, pages 360–367, 1995.

[10] I. A. Essa and A. Pentland. Coding, analysis, interpretation and recognition of facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7), July 1997.

[11] B. Fasel and J. Luettin. Automatic Facial Expression Analysis: A Survey. *Pattern Recognition*, 36(1):259–275, 2003.

[12] I. Fasel, B. Fortenberry, and J. Movellan. A generative framework for real time object detection and classification. *Computer Vision and Image Understanding*, 98:182–210, 2005.

[13] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. In *Proceedings of the Second European Conference on Computational Learning Theory*, 1995.

[14] H. Hong, H. Neven, and C. von der Malsburg. Online facial expression recognition based on personal galleries. In *Proceedings of the International Conference on Automatic Face and Gesture Recognition*, 1998.

[15] M. Isard and A. Blake. Condensation – conditional density propogation for visual tracking. *International Journal of Computer Vision*, 29(1):5–28, 1998.

[16] Q. Ji and X. Yang. Real time visual cues extraction for monitoring driver vigilance. *Lecture Notes in Computer Science: Computer Vision Systems*, 2095:107–124, 2003.

[17] T. Kanade, J. F. Cohn, and Y. Tian. Comprehensive database for facial expression analysis. In *Proceedings of the 4th IEEE International Conference on Automatic Face and Gesture Recognition*, pages 46–53, March 2000.

[18] J.-J. Lien, T. Kanade, J. Cohn, and C. Li. Detection, tracking, and classification of action units in facial expression. *Journal of Robotics and Autonomous Systems*, July 1999.

[19] J. Lim and M.-H. Yang. A direct method for modeling nonrigid motion with thin plate spline. In *Proceedings of the IEEE Conference on Computer Vision and Patern Recognition*, pages 1196–1202, 2005.

[20] C. Lisetti and D. Rumelhart. Facial expression recognition using a neural network. In *Proceedings of the 11th International Flairs Conference*. AAAI Press, 1998.

[21] I. Matthews and S. Baker. Active appearance models revisited. *International Journal of Computer Vision*, 60(2):135–164, 2004.

[22] J. M^{c}Call and M. M. Trivedi. Pose invariant affect analysis using thin-plate splines. In *Proceedings of International Conference on Pattern Recognition*, pages 958–964, August 2004.

[23] K. Okuma, A. Taleghani, N. de Freitas, J. J. Little, and D. G. Lowe. A boosted particle filter: Multitarget detection and tracking. *Lecture Notes in Computer Science: Proceedings of the 8th European Conference on Computer Vision*, 3021:28–39, 2004.

[24] P. Phillips, H. Wechsler, J. Huang, and P. Rauss. The FERET database and evaluation procedure for face-recognition algorithms. *Image and Vision Computing*, 16(5):295–306, 1998.

[25] Y. Rui and Y. Chen. Better proposal distributions: Object tracking using unscented particle filter. In *Proceedings of the IEEE Conference on Computer Vision and Patern Recognition*, pages 786–793, 2001.

[26] H. Schneiderman and T. Kanade. Probabilistic modeling of local appearance and spatial relationships for object recognition. In *Proceedings of the IEEE Conference on Computer Vision and Patern Recognition*, pages 45–51, 1998.

[27] C. Su, Y. Zhuang, L. Huang, and F. Wu. A two-step approach to mulitple facial feature tracking: Temporal particle filter and spatial belief propogation. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*, 2004.

[28] Y.-L. Tian, T. Kanade, and J. F. Cohn. *Recognizing action units for facial expression analysis*, pages 32–66. World Scientific Publishing Co., Inc., River Edge, NJ, USA, 2002.

[29] P. Viola and M. Jones. Robust real-time object detection. *International Journal of Computer Vision*, 2002.