Capturing Subtle Facial Motions in 3D Face Tracking

Zhen Wen, Thomas S. Huang
Beckman Institute
University of Illinois at Urbana-Champaign
Urbana, IL 61801
{zhenwen, huang}@ifp.uiuc.edu

Abstract

Facial motions produce not only facial feature points motions, but also subtle appearance changes such as wrinkles and shading changes. These subtle changes are important yet difficult issues for both analysis (tracking) and synthesis (animation). Previous approaches were mostly based on models learned from extensive training appearance examples. However, the space of all possible facial motion appearance is huge. Thus, it is not feasible to collect samples covering all possible variations due to lighting conditions, individualities, and head poses. Therefore, it is difficult to adapt such models to new conditions. In this paper, we present an adaptive technique for analyzing subtle facial appearance changes. We propose a new ratio-image based appearance feature, which is independent of a person's face albedo. This feature is used to track face appearance variations based on exemplars. To adapt the exemplar appearance model to new people and lighting conditions, we develop an online EM-based algorithm. Experiments show that the proposed method improves classification results in a facial expression recognition task, where a variety of people and lighting conditions are involved.

1. Introduction

Many applications in human computer interaction demand accurate human facial motion tracking (analysis) and realistic animation (synthesis). In the past several decades, great progress has been made in this research area. There are mainly two types of approaches: geometric-feature-based methods and appearance-based methods [1, 3, 8, 10, 11, 16, 21, 22]. Geometric-feature-based approaches model the motions of salient facial points to infer the deformation of a facial surface. The deformation of featureless areas is decided by interpolation. However, facial motions also exhibit detailed appearance changes such as wrinkles and creases as well. These details are important visual cues but they are difficult to analyze and synthesize using geometric-feature-based approaches. Appearance-based approaches try to deal with this problem by using all face image pixel. To reduce the high dimensionality of the appearance space, subspace analysis techniques such as Principal Component Analysis (PCA) [14], are used to find low dimensional approximation of the space. To enhance certain features (e.g., edges), face images can be processed by filtering before extracting appearance-based features. However, the space of all face appearance is huge, affected by the variations across different head poses, individuals, lighting, expressions, speech and etc. Thus it is difficult for appearance-based methods to collect enough face appearance data and train a model that works robustly in many different scenarios. In this respect, the geometric-feature-based methods are more robust to large head motions, changes of lighting and are less person-dependent.

To combine the advantages of both approaches, people have been investigating methods of using both geometry (shape) and appearance (texture) in face analysis and synthesis. The Active Appearance Model (AAM) [7] and its variants, apply PCA to model both the shape variations of image patches and their texture variations. They have been shown to be powerful tools for face alignment, recognition, and synthesis. Blanz and Vetter [4] derived similar models for 3D faces modelling. In facial expression classification, Tian et al. [24] and Zhang et al. [28] proposed to train classifiers (e.g. neural networks) using both shape and texture features. The trained classifiers were shown to outperform classifiers using shape or texture features only. In these approaches, some variations of texture are absorbed by shape variation models. However, the potential texture space can still be huge because many other variations are not modelled by shape model. Moreover, little has been done to adapt the learned models to new conditions. As a result, the application of these methods are limited to conditions similar to those of training data.

In this paper, we present a new technique to capture subtle facial motions in 3D non-rigid face tracking. We propose a ratio-image based appearance feature to remove its dependency on people's facial surface albedos. Therefore, the appearance feature is less person-dependent because facial surface albedos usually contain person-dependent de-

Proceedings of the Ninth IEEE International Conference on Computer Vision (ICCV 2003) 2-Volume Set
0-7695-1950-4/03 $17.00 © 2003 IEEE

tails such as facial hair, permanent wrinkles, freckles, scars and etc. This property helps to adapt corresponding appearance models to new people. Based on the proposed appearance feature, face texture variations due to facial motion are modelled using exemplars. To adapt the appearance models for new people and lighting conditions, we develop an online EM-based algorithm. In our experiments, a facial expression classification task is used to evaluate the efficacy of the proposed method. The results show the proposed approach improves classification results and can be adapted to different subjects and lighting conditions.

The remainder of this paper is organized as follows. We describe the related work in Section 2. Then we present our approach from Section 3 to Section 6. Experimental results are presented in Section 7. Finally, we conclude our paper and discuss future work in Section 8.

2. Related Work

Recently, people proposed to use texture to improve 3D face tracking. La Cascia et al. [5] modelled the face with a texture-mapped cylinder. 3D rigid face tracking was formulated as a texture image registration problem. Pighin et al. [18] and Revert et al. [20] estimated facial deformation based on the discrepancy between a target face image and the image synthesized from reference face texture images. A linear combination of a set of reference texture images was used to cope with the texture variations. However, the set of reference texture images should be of the same person and in the same lighting condition. Moreover, they were computationally expensive because all image pixels are used in optimization.

Because the appearance of facial motions has large variations due to many factors, such as poses, people and lighting conditions, it has been a difficult problem to adapt appearance models of facials motions. Recently, Jepson et al. [13] proposed an online appearance model which could be adapted to temporal facial appearance variations. However, only current stable model of facial appearance was learned and the non-rigid facial motions were not interpreted by the model. Liu et al. [16] used the ratio image technique to map one person's facial expression details to other people's faces. One essential property of the ratio image is that it removes dependency on faces' reflectance property.

An analogy of this adaptation problem in speech recognition domain is speaker adaptation. A good survey can be found in [27]. One type of approach, called Maximum Likelihood Linear Regression (MLLR) [12], is to linearly transform the parameters of a speaker-independent model such that the likelihood of the adaptation data of a particular person is maximized.

This paper describes a method for analyzing subtle facial motion in 3D non-rigid face tracking. It is inspired by the advantages of methods combining geometry and appearance, and the recent advances in adaption algorithms. The key contributions include: (1) a new appearance feature which is independent of a face's reflectance property; (2) an online adaption algorithm to progressively adapt the appearance model to new conditions.

3. Framework Overview

3.1. Facial Motion Formulation

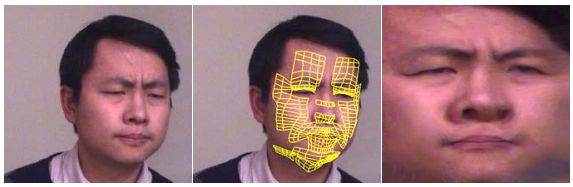
Human face can be modelled as a 3D textured geometric model. An input face image can be aligned to a parametric 3D face surface $S(s,t)$ and warped to a reference plane to extract texture map $I(u,v)$, where (s,t), (u,v) are the coordinate systems of the parametric face surface and the texture plane respectively. The facial motion can be defined as a signal $d(u,v)$ over the texture plane (u,v), which is the 3D displacement of facial surface point at (u,v). Geometric-feature-based face tracking methods sample $d(u,v)$ at the places of facial salient points. Because each salient point needs a textured local support region for robust motion estimation and a human face contains poorly textured area, the number of samples is limited. Thus aliasing is unavoidable. To analyze the subtle motions in $d(u,v)$, we can decompose $d(u,v)$ as

$$d(u,v) = d_L(u,v) + d_H(u,v) \tag{1}$$

where $d_L(u,v)$ is the low frequency component which can be analyzed by geometric-feature-based methods. The residual $d_H(u,v)$ contains subtle facial motion which are important visual cues. In this paper, $d_L(u,v)$ is referred as geometric deformation. Its variation can be modelled using subspace techniques such as PCA [20] or customized "Action Units" [11, 21]. Because it is difficult to measure $d_H(u,v)$ directly, the related texture $I(u,v)$ is usually used for analysis instead. Moreover, we believe $d_H(u,v)$ exhibits larger variation than $d_L(u,v)$ across different individuals and lighting conditions. Thus a good low-dimensional subspace approximation of $d_H(u,v)$ variation may be difficult. Nevertheless, facial motions exhibit common semantic exemplars such as typical expressions and visemes, which makes it meaningful to use exemplar-based approach such as [25].

3.2. Framework Overview

Given a face video, we use a geometric-feature-based method [21] to estimate 3D geometric deformation $d_L(u,v)$. A 3D face mesh model is aligned with the first face image using interactively selected facial points. Facial geometric deformation is approximated by a linear combination of 12 "Action Units" (AUs): $\vec{V} = \mathbf{L}\vec{p}$, where $\vec{V}$ is



(a) (b) (c)

Figure 1: *(a): input video frame; (b): snapshot of the geometric tracking system. (c): extracted texture map*

the geometric deformation, $\mathbf{L}$ is a matrix containing AU basis vectors, and $\vec{p}$ is the coefficient vector of the AUs. $\vec{p}$ is used as the geometric feature for facial deformation. The facial deformation observed in image plane can be represented by $\vec{V}_{2D} = \mathbf{P}(\mathbf{R}(\vec{V}_0 + \mathbf{L}\vec{p}) + \vec{T})$, where $\mathbf{P}$ is the projection matrix, $\mathbf{R}$ is the 3D rotation matrix, and $\vec{T}$ is the translation. The camera parameters in $\mathbf{P}$ are known in our implementation. The inter-frame projected facial motion can be approximated by the derivative of $\vec{V}_{2D}$ with respect to the unknown parameters of $\mathbf{R}$, $\vec{T}$ and $\vec{p}$. The derivative is linearized as in [21], which can be solved by least square method. We estimate the inter-frame motion using template-matching based optical flow. Fig. 1(b) shows a snapshot of the geometric tracking system, where a yellow mesh is used to visualize the geometric motions of the face. The input video frame is shown in Fig 1(a).

The geometric deformation parameters determine the registration of each image frame to the face texture map. Thus we can derive a sequence of face texture maps, which are independent of geometric motion $d_L(u, v)$. Fig. 1(c) shows the extracted texture map. From the texture maps, we extract appearance-based features described in section 4. These features are designed for subtle details of facial expression and independent of people's face surface albeo. We then use the appearance features, together with shape features, to analyze face appearance variations based on semantically meaningful exemplars.

To extend a trained appearance-based exemplar model to new conditions, an online EM-based algorithm is used to update the exemplar model progressively.

4. Ratio-image-based Appearance Feature

4.1. Ratio image

We assume faces are Lambertian. Let $\vec{n}(u, v)$, $\rho(u, v)$ denote the normal and albedo of a face surface point at texture plane (u, v). Let $\mathcal{L}$ denote the distant lighting distribution. The irradiance on the face is then a function of $\vec{n}(u, v)$, given by an integral over the upper hemisphere $\Omega(\vec{n}(u, v))$ at $\vec{n}(u, v)$.

$$E(u, v) = \int_{\Omega(\vec{n}(u,v))} \mathcal{L}(\omega)(\vec{n}(u, v) \cdot \omega) d\omega \qquad (2)$$

The intensity of the neutral face point p at (u, v) is $I(u, v) = \rho(u, v)E(u, v)$. After the face surface is deformed, the intensity of p is $I'(u, v) = \rho(u, v)E'(u, v)$. We denote

$$\Re(u, v) = \frac{I'(u, v)}{I(u, v)} = \frac{E'(u, v)}{E(u, v)} \qquad (3)$$

It can be observed that $\Re(u, v)$, called the ratio image, is independent of surface reflectance property $\rho(u, v)$ [16]. Therefore, $\Re(u, v)$ can be used to characterize facial motions of faces with different albedos.

4.2. Feature extraction

To use $\Re(u, v)$ in face tracking, more compact features need to be extracted from the high dimensional ratio image. First, as explained by Section 3.1, low frequency variation of facial motion could be captured by geometric-feature-based methods. Thus we extract features from $\Re(u, v)$ in frequency domain and use the high frequency components as the features for $d_H(u, v)$. Second, past studies on facial motions [28, 24] have shown that there are certain facial areas where high frequency appearance changes are more likely to occur and thus suitable for texture feature extraction. We apply this domain knowledge in our feature extraction. However, because of noise in tracking and individual variation, it is difficult to locate these locations automatically with enough precision. Therefore, we extract the texture-based features in facial regions instead of points, and then use the weighted average as the final feature. Eleven regions are defined on the geometric-motion-free texture map. These eleven regions are highlighted on the texture map in Fig. 2. Note that these regions can be considered constant in the automatically extracted texture map, where the facial feature points are aligned by geometric tracking.

Figure 2: *Selected facial regions for feature extraction*

Gabor wavelets are used to extract the appearance changes as a set of multi-scale and multi-orientation coefficients. In our implementation, we use two spatial frequency scales with wavelength of 5 and 8 pixels, and 6 orientations

at each scale. Thus for each point, we have $2 \times 6 = 12$ Gabor wavelets coefficients. We choose to compute the Gabor wavelets coefficients of the logarithm of $\Re(u, v)$, denoted by $Z(u, v)$. Based on equation (3) and the linearity property of Gabor transform, we have

$$\begin{aligned} Z(u,v) &= G(\log(\Re(u,v))) \\ &= G(\log(I'(u,v))) - G(\log(I(u,v))) \end{aligned} \quad (4)$$

where function G denotes a Gabor transform as in [24, 28]. We impose a positive lower bound on pixel values in texture I' and I to avoid singular situations. In our approach, only the magnitudes of Gabor transform results are used because the phases are very sensitive to noise in positions. Then we note that if $Z(u, v) < 0$, it means the neutral face texture I contains more high frequency components than the deformed face texture I'. It could be caused by one of the following reasons: (1) the misalignment of I' and I; (2) high gradient of $\log(I)$ due to low intensities of I; (3) flattening of wrinkles and creases on neutral face during motion. Scenarios (1) and (2) should be considered as noise, and (3) rarely happens in common human facial motions. Thus we discard negative values of $Z(u, v)$.

In practice, we need to account for the foreshortening effect of the texture projection. For a 3D face surface patch, the larger its visible area in input image, the higher confidence we should have on the extracted features of the corresponding texture patch. To this end, we construct a confidence map $\kappa(u, v)$ following [5], which is based on the ratio of each 3D surface patch's projected area in the texture plane and its area in the input image. For each facial motion region q ($q = 1...11$), we compute a confidence coefficient c_q as the average of the $\kappa(u, v)$ in this region. The resulting confidence coefficients are used to weight the features in tracking described in Section 5.

$\Re(u, v)$ contains noise due to misalignment of I' and I. To reduce the influences of noise on the appearance feature, we construct another weight map $w(u, v)$, which tries to give large weight for features in deformed area and small weight for features in un-deformed area. We define $w(u, v) = 1 - corr(u, v)$ in similar spirit as [16], where $corr(u, v)$ is the normalized cross correlation coefficient between two patches centered at (u, v) from $G(\log(I'))$ and $G(\log(I))$. The idea is that high frequency components of $\log(I')$ and $\log(I)$ should be close for un-deformed area, since I' and I are roughly aligned by geometric-feature-based tracking. We use $w(u, v)$ to compute the weighted average of Gabor wavelets coefficients in the 11 selected regions, resulting 11 appearance feature vectors of 12-dimension.

4.3. Influences of lighting

Under the assumptions of Lambertian faces, distant illumination and ignoring cast shadows, the proposed appearance features are not sensitive to changes of lighting conditions. According to [2, 19], the irradiance can be represented by a linear combination of spherical harmonic basis function. For Lambertian surfaces, only the first 2 orders of the basis functions (9 basis) are needed to approximate the irradiance, that is

$$E(u,v) \approx \sum_{l \le 2, -l \le m \le l} \hat{A}_l \, L_{lm} \, Y_{lm}(\vec{n}(u,v)) \quad (5)$$

where $\hat{A}_l$ is a constant, L_{lm} is a coefficient decided by lighting, and Y_{lm} is the spherical harmonic basis function. Assuming that neutral face and the deformed face are in the same lighting condition, we have

$$\begin{aligned} \Re(u,v) &= \frac{E'(u,v)}{E(u,v)} \\ &\approx \frac{\sum_{l \le 2, -l \le m \le l} \hat{A}_l \, L_{lm} \, Y_{lm}(\vec{n'}(u,v))}{\sum_{l \le 2, -l \le m \le l} \hat{A}_l \, L_{lm} \, Y_{lm}(\vec{n}(u,v))} \end{aligned} \quad (6)$$

The high frequency facial motion $d_H(u, v)$ can produce high frequency differences between $\vec{n'}(u, v)$ and $\vec{n}(u, v)$, and therefore between $Y_{lm}(\vec{n'}(u, v))$ and $Y_{lm}(\vec{n}(u, v))$. The irradiance $E'(u, v)$ will contain a linear combination of these high frequency differences weighted by the lighting coefficients. In other words, high frequency changes in $\Re(u, v)$ are due to $d_H(u, v)$, while lighting will only modulate them in low frequency. If the neutral face and the deformed face are in different lighting conditions, the neutral face texture can be relit to the lighting of the deformed face using face relighting technique in [26].

5. Exemplar-based Texture Tracking

To track the face appearance variation, we choose exemplars set $\Xi = \{x_k, k = 1, ..., K\}$, which are semantically meaningful such as face expressions or visemes. A texture image is interpreted as a state variable X of the exemplars. Unlike [25], these exemplars incur both shape and texture changes. Let Y_S and Y_T denote the shape and texture features respectively. The observation is $Y = \{Y_S, Y_T\}$. We assume Y_S and Y_T are conditionally independent given X. The observation likelihood is

$$p(Y|X) = p(Y_S, Y_T|X) = p(Y_S|X)\, p(Y_T|X) \quad (7)$$

In addition, we assume the texture features in different facial motion regions are independent given X. Their log likelihoods are weighted by confidence coefficients c_q to account for foreshortening effect. That is

$$\log p(Y_T|X) = \sum_{q=1}^{Q} c_q \, \log p(Y_{T_q}|X) \quad (8)$$

where Q is the number of facial motion regions ($Q = 11$). $p(Y_S|X)$ and $p(Y_{T_q}|X)$ are modelled using Gaussian Mixture Model (GMM), assuming diagonal covariance matrices. The feature vectors are normalized by their magnitudes. If the neutral face is chosen as an exemplar, we assign the likelihood using a neutral face classifier such as [23].

In this paper, we do not address the modelling of exemplar dynamics, i.e., we assume uniform conditional density $p(X_t|X_{t-1})$. Techniques for modelling dynamics can be found in [25]. Assuming uniform priors, we have $p_t(X_t) \propto p(Y_{S_t}, Y_{T_t}|X_t)$. The exemplar tracking result can be displayed as $\hat{X}_t = \arg\max p_t(X_t)$.

6. Online EM-based Adaptation

A trained model for facial motion exemplars may work poorly if it can not adapt to lighting changes, or differences in a new individual's exemplars. Fast adaptation algorithm is needed to avoid re-training the model from scratch. Furthermore, it is tedious to collect and label new training data for each new condition. Therefore, we propose to progressively update the model during tracking in an unsupervised way. Because the geometric features are less person-dependent and less sensitive to lighting changes, we assume the geometric component of the initial exemplar model can help to "confidently" track some new data samples. Then the Expectation-Maximization (EM) framework [9] can be applied to update the model parameters. At time t, the E-step provides exemplar ownership probabilities defined as

$$o_{k,t}(Y_t) = \frac{p(Y_t|X_t = k)}{\sum_{k=1}^{K} p(Y_t|X_t = k)} \tag{9}$$

where k is the index of the exemplars. In the M-step, the model is adapted by computing new maximum likelihood estimates of its parameters. Note that we only adapt the texture part of the model because shape features are less person-dependent and not sensitive to changes of lighting.

The idea of Maximum Likelihood Linear Regression (MLLR) can be generalized to this adaptation problem, where we estimate a linear transformation of the GMM mean vectors to maximize the likelihood of new observations. However, conventional MLLR is not an online method which requires multiple data samples for maximum likelihood optimization. In the M-step of our online EM algorithm, only one data sample is available at a time. Thus we restrict the transformation of the GMM mean vectors to be translation only. The M-step of our algorithm is then to estimate $\Delta\mu_{q,k,t}$, which denotes the translation of the GMM mean vectors from initial model, for the q^{th} facial motion region, the k^{th} exemplar at time t. To weight the current data sample appropriately against history, we consider the data samples under an exponential envelope located at the current time as in [13], $F_t(j) = \alpha e^{-(t-j)/\tau}$, for $j \leq t$. Here, $\alpha = 1 - e^{-1/\tau}$.

For the GMM model of certain q, k, t value, suppose the GMM has M components: $\{N(\mu_1, \Sigma_1), ..., N(\mu_M, \Sigma_M)\}$. Here the q, k, t subscripts are dropped for simplicity. Given an adaptation data sample $Y = \{Y_S, Y_T\}$, the ML estimate of the translation $\Delta\mu$ can be computed by solving equation (10) according to [12]:

$$\sum_{m=1}^{M} \gamma_m \Sigma_m^{-1} Y_T \mu_m^T = \sum_{m=1}^{M} \gamma_m \Sigma_m^{-1} (\mu_m + \Delta\mu) \mu_m^T \tag{10}$$

where γ_m is GMM component occupancy probability defined as the probability that Y_T draws from the m^{th} component of the GMM given Y_T draws from this GMM. Y_T is the texture feature of the current adaptation data. A closed-form solution for equation (10) is feasible when Σ_m is diagonal. The i^{th} element of $\Delta\mu$ can be computed as

$$\Delta\mu_i = \frac{\sum_{m=1}^{M} \gamma_m \sigma_{m,i} Y_{T_i} - \sum_{m=1}^{M} \gamma_m \sigma_{m,i} \mu_{m,i}}{\sum_{m=1}^{M} \gamma_m \sigma_{m,i}} \tag{11}$$

where $\sigma_{m,i}$ is the i^{th} diagonal element of Σ_m^{-1}, and Y_{T_i} is the i^{th} element of Y_T.

The probability weighted average of the translation vector up to time t, for the k^{th} exemplar of a certain facial motion region is then

$$\bar{\Delta\mu}_{k,t} = \frac{1}{\beta_{k,t}} \left(\sum_{j=-\infty}^{t} F_t(j) o_{k,t}(Y_j) \Delta\mu_{k,j} \right) \tag{12}$$

where $\beta_{k,t}$ is a normalization factor defined as $\beta_{k,t} = \sum_{j=-\infty}^{t} F_t(j) o_{k,t}(Y_j)$, and $\Delta\mu_{k,j}$ is computed using equation (11). Equation (12) can be rewritten in a recursive manner as:

$$\bar{\Delta\mu}_{k,t} = \frac{(1-\alpha)\beta_{k,t-1} \bar{\Delta\mu}_{k,t-1} + \alpha\, o_{k,t}(Y_t) \Delta\mu_{k,t}}{\beta_{k,t}} \tag{13}$$

where $\beta_{k,t} = (1-\alpha)\beta_{k,t-1} + \alpha\, o_{k,t}(Y_t)$. Equations (11) and (13) are applied to updated the translation vectors of each facial motion region.

In an online EM algorithm, the training samples can not be used iteratively for optimization. Thus, errors made by the initial model may cause the adaptation method to be unstable. Note that the chosen exemplars are common semantic symbols which have their intrinsic structure. Therefore, when adapting the exemplar model to a particular person, the translations of the mean vectors should be constrained rather than random. For the exemplars model of a facial motion region at time t, let ξ denote the translation vector constructed by concatenating the translations of all the mean vectors, i.e., $\xi = [\Delta\mu_1^T ... \Delta\mu_K^T]^T$, where K is the number of exemplars. We impose the constraint that the

vector ξ should lie in certain low dimensional subspaces. To learn such a low dimensional subspace, we first learn a person-independent exemplar model from exemplars of many people. Then a person-dependent exemplar model is learned for each person. We collect a training sample set $\{\xi\}$, consisting of the translation vectors between the mean vectors of person-independent model and person-dependent models. Finally, PCA is applied on the set $\{\xi\}$. Principal orthogonal components which account for major variation in set $\{\xi\}$ are chosen to span a low dimensional subspace. A translation vector can be projected to the learned subspace by

$$a = W^T (\xi - \bar{\xi}) \tag{14}$$

and the new constrained translation vector can be reconstructed as

$$\hat{\xi} = W\, a + \bar{\xi} \tag{15}$$

where W is the matrix consisting of principal components, $\bar{\xi}$ is the mean translation vector of the vectors in $\{\xi\}$.

In summary, the online EM-based adaptation algorithm is as follows:

- E-step: compute exemplar ownership probabilities $o_{k,t}(Y_t)$ based on equations (7), (8) and (9).
- M-step: for the q^{th} facial motion region, estimate the translator vector $\Delta\mu_{q,k,t}$ based on equations (11) and (13). Then we construct the vector $\xi_{q,t}$ and project it using equation (14). Finally, the constrained estimate of $\xi_{q,t}$ is given by equation (15).

7. Experimental Results

We evaluate the efficacy of the proposed tracking method by using the extracted features in a facial expression classifications task. The public available CMU Cohn-Kanade expression database [15] is used. From the database, we selected 47 subjects who has at least 4 coded expression sequences. Overall the selected database contains 2981 frames. There are 72% female, 28% male, 89% Euro-American, 9% Afro-American, and 2% Asian. Several different lighting conditions are present in the selected database. The image size of all the data is 640×480. For the Cohn-Kanade database, Tian and Bolle [23] achieved a high neutral face detection rate using geometric features only. That indicates the database does not contain expressions with little geometric motion yet large texture variation. Using geometric feature only on the database, Cohen et al. [6] reported good recognition results for happiness and surprise, but much more confusion among anger, disgust, fear and sadness. In this section, we present our experimental results showing the proposed method improves the performance for these four expressions.

We select seven exemplars including six expressions and *neutral*. The six expressions are anger, disgust, fear, happiness, sadness, and surprise. In our experiments, we first assign *neutral* vs. non-*neutral* probability using a neutral network similar to [23], which achieved a recognition rate of 92.8% for *neutral*. For the remaining exemplars, we use 4 components for each GMM model. The tracking results are used to perform facial expression classification as $\hat{X}_t = \arg\max p_t(X_t)$. Although this classifier may not be as good as more sophisticated classifiers such as those in [1, 6, 10, 28], it can be used as a test-bed to measure the relative performances of different features and the proposed adaptation algorithm.

In the first experiment, we compare the classification performances of using geometric feature only and using both geometric and ratio-image-based appearance features. We use 60% data of each person as training data and the rest as test data. Thus it is a person-dependent test. The recognition rates for the four easily confused expressions are shown in table 1. It can be observed that appearance features significantly improve results. For happiness and surprise, the results of the two methods are close. It is consistent with [6]. We choose to omit them due to space limit.

Expressions	Anger	Disgust	Fear	Sadness
Geo-only	74.8%	76.5%	65.6%	77.1%
Proposed	92.7%	85.7%	81.5%	90.5%

Table 1: *Comparison of the proposed approach with geometric-only method in person-dependent test.*

In the second experiment, we compare the classification performances of ratio-image based appearance feature and non-ratio-image based appearance feature. The non-ratio-image based feature does not consider the neutral face texture, and is computed as $G(\log(I'(u,v)))$ instead of using equation (4). To show the advantage of ratio-image based feature in the ability to generalize to new people, the test is done in a person-independent way. That is, all data of one person is used as test data and the rest as training data. This test is repeated 47 times, each time leaving a different person out (leave one out cross validation). The person-independent test is more challenging because the variations between subjects are much larger than those within the same subject. To factor out the influence of geometric feature, only the appearance feature is used for recognition in this experiment. The average recognition rates are shown in table 2. We can see that ratio-image based feature outperforms non-ratio-image based feature significantly. For individual subject, we found that the results of the two features are close when the texture does not have much details. Otherwise, ratio-image based feature is much better.

The third experiment again uses person-independent set-

Expressions	Anger	Disgust	Fear	Sadness
ratio	37.0%	59.6%	35.7%	41.8%
non-ratio	24.7%	22.1%	24.4%	15.6%

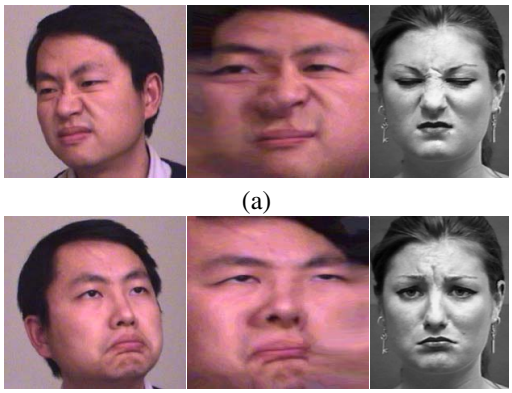
Table 2: *Comparison of the proposed appearance feature (ratio) with non-ratio-image based appearance feature (non-ratio) in person-independent recognition test.*

ting and leave one out cross validation. For each test, we use 50% of the data of the test person as adaptation data and the rest as test data. Without applying adaptation algorithm, we first compare the performances of using geometric feature only and using both geometric and ratio-image-based appearance features. The results are shown in the rows (a) and (b) of table 3. It can be observed that improvement is less significant than that in the first experiment. This is mainly due to the individual variations in facial expressions. Then, We test the performance of the proposed online EM-based adaptation algorithm. Only the models of the four easily confused expressions are adapted. In each test, we apply PCA to the training data. The first 11 principal components are selected which account for about 90% of total variations. The adaptation is online and unsupervised, without using the labels of the adaptation data. We choose $\alpha = 0.1$ for fast adaptation because the amount of the adaptation data is limited. The recognition rates the adaptation are shown in the row (d) of table 3. We can see the adaption algorithm improves the recognition rates. For comparison, we also show in the row (c) the recognition rates of adaptation without the PCA subspace constraints. It can be seen that the unconstrained adaptation can make the performance worse.

Expressions	Anger	Disgust	Fear	Sadness
a	66.6%	65.3%	60.8%	69.8%
b	70.7%	70.2%	64.6%	72.5%
c	75.3%	59.2%	64.0%	73.1%
d	77.9%	78.1%	67.7%	77.8%

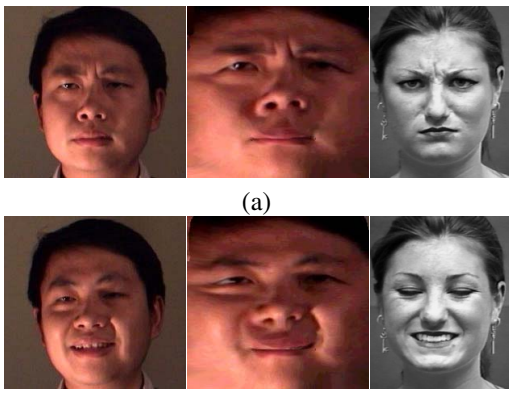
Table 3: *Comparison of different algorithms in person-independent recognition test. (a): using geometric feature only, (b:) using both geometric and ratio-image based appearance feature, (c): applying unconstrained adaptation, (d) applying constrained adaptation.*

To test the proposed method under large 3D rigid motions and novel lighting conditions, we also collect two video sequences of a subject who is not in the training database. The frame size of the videos is 640×480. The first video has 763 frames and contains substantial global face translation and rotation. The second video has 435 frames and is taken under a lighting condition dramatically different from the rest of data. We manually label the image frames using the seven categories as the ground truth. Two snapshots for each sequence are shown in Fig. 3 and 4. The corresponding recognition results are also illustrated using one of the training example. We compare the expression recognition rates of the proposed method with geometric-feature-only method. The overall average recognition rate of our method is 71%, while the rate of the geometric-only method is 59%. Part of the tracking results is visualized in the accompanying videos. In the videos, the upper left of the frame is the input video frame, the upper right is the geometric feature based tracking visualized by a yellow mesh. The exemplar $\hat{X}_t$ is shown on the bottom. We can observe that our method can still track the texture variations when there are large 3D motions, or under dramatically different lighting conditions.



(a) (b)

Figure 3: *The results under different 3D poses. For both (a) and (b): Left: cropped input frame. Middle: extracted texture map. Right: recognized expression.*



(a) (b)

Figure 4: *The results in a different lighting condition. For both (a) and (b): Left: cropped input frame. Middle: extracted texture map. Right: recognized expression.*

8. Conclusions and Future Work

We have presented a method for analyzing subtle facial motion in 3D face tracking. We propose a ratio-image based

feature for appearance variation due to subtle facial motions. Because this feature is independent of people's face albedos, it helps appearance model adapt to new people. Using the proposed feature extracted from stabilized face texture, we analyze subtle facial motion based on exemplars. An online EM-based algorithm is developed to adapt the model parameters in new conditions. We demonstrated the efficacy of our method in facial expression classification experiments where a variety of people and lighting conditions are involved.

We are planning to incorporate dynamics models in our method following [25]. Dynamics model would improve the temporal behavior of our method and the capability of the adaptation algorithm. The correlation between shape and texture features will be investigated for better fusion method. We are also planning to extend our method to model the appearance of oral cavity, which is important for lip-reading applications. Additional appearance features, such as those based on DCT [17] may be incorporated to model the novel texture inside the mouth.

Acknowledgments

This work was supported in part by National Science Foundation Grants CDA 96-24386 and IIS-00-85980. We would like to thank Jilin Tu for help in the algorithm implementation. We also thank Zicheng Liu, Zhengyou Zhang, and anonymous reviewers for their valuable comments.

References

[1] M. Bartlett, J. Hager, P. Ekman, and T. Sejnowski. Measuring facial expressions by computer image analysis. *Psychophysiology*, pages 253–263, 1999.

[2] R. Basri and D. Jacobs. Lambertian reflectance and linear subspaces. In *ICCV*, pages 383–390, 2001.

[3] M. J. Black and Y. Yacoob. Tracking and recognizing rigid and non-rigid facial motions using local parametric models of image motions. In *ICCV*, pages 374–381, 1995.

[4] V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. In *SIGGRAPH 99*, pages 187–194, 1999.

[5] M. L. Cascia, S. Sclaroff, and V. Athitsos. Fast, reliable head tracking under varying illumination: An approach based on robust registration of texture-mapped 3d models. *IEEE PAMI*, 22(4):322–336, 2000.

[6] I. Cohen and et al. Facial expression recognition from video sequences: Temporal and static modeling. *CVIU*, 2003.

[7] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. In *ECCV*, pages 484–498, 1998.

[8] D. DeCarlo and D. Metaxas. Optical flow constraints on deformable models with applications to face tracking. *IJCV*, 38(2):99–127, 2000.

[9] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the em algorithm. *J. Royal Stat. Soc. B*, 39:1–38, 1977.

[10] G. Donato and et al. Classifying facial actions. *IEEE PAMI*, 21(10):974–989, 1999.

[11] I. Essa and A. Pentland. Coding analysis, interpretation, and recognition of facial expressions. *IEEE PAMI*, pages 757–763, 1997.

[12] M. Gales and P. Woodland. Mean and variance adaptation within the mllr framework. *Computer Speech and Language*, 10:249–264, 1996.

[13] A. Jepson, D. Fleet, and T. El-Maraghi. Robust, on-line appearance models for vision tracking. In *CVPR*, pages 415–422, 2001.

[14] I. T. Jolliffe. *Principal Component Analysis*. Springer-Verlag, 1986.

[15] T. Kanade, J. Cohn, and Y. Tian. Comprehensive database for facial expression analysis. In *Proc. of Int'l Conf. on Automated Face and Gesture Recognition*, pages 45–63, 2000.

[16] Z. Liu, Y. Shan, and Z. Zhang. Expressive expression mapping with ratio images. In *SIGGRAPH*, pages 271–276, 2001.

[17] I. Matthews and et al. A comparison of model and transform-based visual features for audio-visual lvcsr. In *ICME*, 2001.

[18] F. Pighin, D. H. Salesin, and R. Szeliski. Resynthesizing facial animation through 3d model-based tracking. In *ICCV*, pages 143–150, 1999.

[19] R. Ramamoorthi and P. Hanrahan. A signal-processing framework for inverse rendering. In *SIGGRAPH*, pages 117–128, 2001.

[20] L. Reveret and I. Essa. Visual coding and tracking of speech related facial motion. In *Proc. of Workshop on Cues in Communication*, 2001.

[21] H. Tao and T. S. Huang. Explanation-based facial motion tracking using a piecewise bezier volume deformation model. In *CVPR*, 1999.

[22] D. Terzopoulos and K. Waters. Analysis of dynamic facial images using physical and anatomical models. In *ICCV*, pages 727–732, 1990.

[23] Y. Tian and R. M. Bolle. Automatic neutral face detection using location and shape features. Computer Science Research Report RC 22259, IBM Research, 2001.

[24] Y. Tian, T. Kanade, and J. Cohn. Evaluation of gabor-wavelet-based facial action unit recognition in image sequences of increasing complexity. In *Proc. of Int'l Conf. on Automated Face and Gesture Recognition*, 2002.

[25] K. Toyama and A. Blake. Probabilistic tracking with exemplar in a metric space. *IJCV*, 48(1):9–19, 2002.

[26] Z. Wen, Z. Liu, and T. S. Huang. Face relighting with radiance environment maps. In *CVPR*, volume 2, pages 158–165, 2003.

[27] P. Woodland. Speaker adaptation: Techniques and challenges. In *Int'l Workshop on Automatic Speech Recognition and Understanding*, 1999.

[28] Z. Zhang and et al. Comparison between geometric-based and gabor-wavelets-based facial expression recognition using multi-layer perceptron. In *Proc. of Int'l Conf. on Automatic Face and Gesture Recognition*, pages 454–459, 1998.