Facial Expression Recognition Based on 3D Dynamic Range Model Sequences

Yi Sun and Lijun Yin

Department of Computer Science, State University of New York at Binghamton
Binghamton, New York, 13902 USA

Abstract. Traditionally, facial expression recognition (FER) issues have been studied mostly based on modalities of 2D images, 2D videos, and 3D static models. In this paper, we propose a spatio-temporal expression analysis approach based on a new modality, 3D dynamic geometric facial model sequences, to tackle the FER problems. Our approach integrates a 3D facial surface descriptor and Hidden Markov Models (HMM) to recognize facial expressions. To study the dynamics of 3D dynamic models for FER, we investigated three types of HMMs: temporal 1D-HMM, pseudo 2D-HMM (a combination of a spatial HMM and a temporal HMM), and real 2D-HMM. We also created a new dynamic 3D facial expression database for the research community. The results show that our approach achieves a 90.44% person-independent recognition rate for distinguishing six prototypic facial expressions. The advantage of our method is demonstrated as compared to methods based on 2D texture images, 2D/3D Motion Units, and 3D static range models. Further experimental evaluations also verify the benefits of our approach with respect to partial facial surface occlusion, expression intensity changes, and 3D model resolution variations.

1 Introduction

Research on FER has been based primarily on findings from Psychology and particularly on the Facial Action Coding System [1]. Many successful approaches have utilized Action Units (AU) recognition [2,3,4,5,6,7,8] or Motion Units (MU) detection [9,10,11]. Other well-developed approaches concentrate on facial region features, such as manifold features [12] and facial texture features [13,14]. Ultimately, however, all of above methods focus on most commonly used modality: 2D static images or 2D videos.

Recently, the use of 3D facial data for FER has attracted attention as the 3D data provides fine geometric information invariant to pose and illumination changes. There is some existing work for FER using 3D models created from 2D images [15] or from 3D stereo range imaging systems [16,17]. However, the 3D models that have been used are all *static*. The most recent technological advances in 3D imaging allow for real-time 3D facial shape acquisition [18,19] and analysis [20]. Such 3D sequential data captures the dynamics of time-varying facial surfaces, thus allowing us to use 3D dynamic surface features or 3D motion units (rather than 2D motion units) to scrutinize facial behaviors at a detailed level. Wang *et al* [18] have successfully developed a hierarchical framework for tracking high-density 3D facial sequences. The recent work in [20] utilized dynamic 3D models of six subjects for facial analysis and editing based on the generalized facial manifold of a standard model.

D. Forsyth, P. Torr, and A. Zisserman (Eds.): ECCV 2008, Part II, LNCS 5303, pp. 58–71, 2008.

Motivated by the recent work of 3D facial expression recognition reported by Yin *et al* [16] based on a static 3D facial expression database [21], we extend the facial expression analysis to a dynamic 3D space. In this paper, we propose a spatio-temporal 3D facial expression analysis approach for FER using our newly-created 3D dynamic facial expression database. This database contains 606 3D facial video sequences with 101 subjects: each subject has six 3D sequences corresponding to six prototypic facial expressions. Our approach uses 3D labeled surface type to represent the human facial surface and transforms the feature to an optimal compact space using linear discriminative analysis. Such a 3D surface feature representation is relatively robust to changes of pose and expression intensities. To explore the dynamics of 3D facial surfaces, we investigated a 1D temporal HMM structure and extended it to a pseudo-2D HMM and a real 2D HMM. There have been existing HMM-based approaches for FER using 2D videos [7,9,22], by which either a 1D HMM or multi-stage 1D-HMM was developed. However, no true 2D-HMM structure was applied to address the FER problem. Our comparison study shows that the proposed real 2D-HMM structure is better than the 1D-HMM and pseudo 2D-HMM in describing the 3D spatio-temporal facial properties.

In this paper, we conducted comparative experiments using our spatio-temporal 3D model-based approach with approaches based on 2D/3D motion units, 2D textures, and 3D static models. The experimental results show that our approach achieves a 90.44% person-independent recognition rate in distinguishing the six prototypic expressions, which outperforms the other compared approaches. Finally, the performance of our proposed approach was evaluated on its robustness dealing with 1) partial facial surface occlusion, 2) expression intensity changes, and 3) 3D model resolution variations. The paper is organized as follows: we first introduce our new 3D dynamic facial expression database in Section 2. We then describe our 3D facial surface descriptor in Section 3 and the HMM classifiers in Section 4. The experimental results and analysis are reported in Section 5, followed by the conclusion in Section 6.

2 Dynamic 3D Face Database

There are some existing public 3D *static* face databases, such as FRGC 2.0 [23], BU-3DFE [21], etc. However, to the best of our knowledge, there is no *3D dynamic* facial expression database publicly available. To investigate the usability and performance of the 3D dynamic facial models for FER, we created a dynamic 3D facial expression database [24] using the Dimensional Imaging's 3D dynamic capturing system [19]. The system captures a sequence of stereo images and produces the range models using a passive stereo-photogrammetry approach. At the same time, 2D texture videos of the dynamic 3D models are also recorded. Figure 1 shows the dynamic 3D face capture system with three cameras. Each subject was requested to perform the six prototypic expressions (*i.e., anger, disgust, fear, smile, sad, and surprise*) separately. Each 3D video sequence captured one expression at a rate of 25 frames per second and each 3D video clip lasts approximately 4 seconds with about 35,000 vertices per model. Our database currently consists of 101 subjects including 606 3D model sequences with 6 prototypic expressions and a variety of ethnic/racial ancestries. An example of a 3D facial sequence is shown in Figure 1. More details can be found in [24].

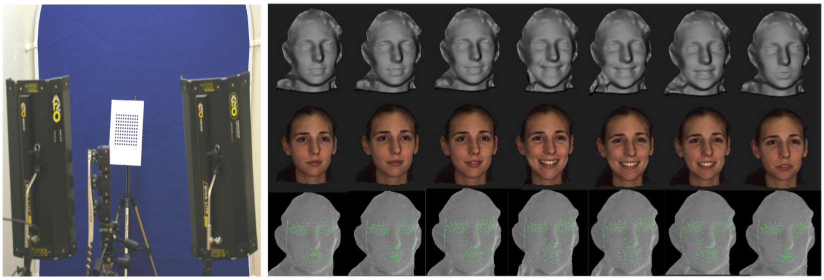


Fig. 1. Left:Dynamic 3D face capturing system setup. Right: sample videos of a subject with smile expression(from top to bottom: shaded models, textured models, and wire-frame models with 83 tracked control points).

3 3D Dynamic Facial Surface Descriptor

The dynamic 3D face data provides both facial surface and motion information. Considering the representation of facial surface and the dynamic property of facial expressions, we propose to integrate a facial surface descriptor and Hidden Markov Models to analyze the spatio-temporal facial dynamics. It is worth noting that we aim at verifying the usefulness and merits of such 3D dynamic data for FER in contrast to the 2D static/dynamic data or 3D static data. Therefore, we do not focus on developing a fully automatic system for FER in this paper. Our system is outlined in Figure 2, which consists of model pre-processing, HMM-based training, and recognition.

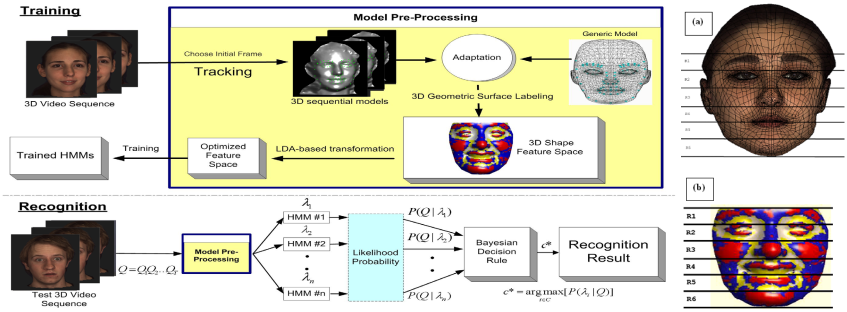


Fig. 2. Left: Framework of the FER system. Right: sub-regions defined on an adapted model (a) and a labeled model (b).

In the first stage, we adapt a generic model (*i.e., tracking model*) to each range model of a 3D model sequence. The adaptation is controlled by a set of 83 pre-defined key points (colored points on the generic model in Figure 2). After adaptation, the correspondence of the points across the 3D range model sequence is established. We apply a surface labeling approach [25] to assign each vertex one of eight primitive shape types. Thus, each range model in the sequence is represented by a “label map”, *G*, as shown in the 3D shape feature space of Figure 2, where different colors represent different labeled shape types. We use Linear Discriminative Analysis (LDA) to transform the label map to an optimal compact space to better separate different expressions. Given the optimized features, the second stage is to learn one HMM for each expression. In recognition, the temporal/spatial dynamics of a test video is analyzed by the trained HMMs. As a result, the probability scores of the test video to each HMM are evaluated by the Bayesian decision rule to determine the expression type of the test video.

3.1 Facial Model Tracking and Adaptation

As the high-resolution range models vary in the number of vertices across 3D video frames, we must establish the vertices' correspondences and construct a common feature vector. To do so, we applied a generic model adaptation approach to "sample" the range models. This process consists of two steps: control points tracking and generic model adaptation. A set of 83 pre-defined key points is tracked using an active appearance model based approach on 2D video sequences [26,19], where the key points in the initial frame are manually picked. To reduce the tracking error, a post-processing procedure was applied by manually correcting some inaccurately tracked points. Since the 2D texture and the 3D range model of each frame are matched accurately from the system, the key points tracked in the 2D video can be exactly mapped to the 3D range surface. This semi-automatic approach allows us to obtain accurate control points on the sequential models. Figure 1 (bottom row) shows an example of a tracked sequence. The adaptation procedure is as follows: Given the N (=83) control points $\boldsymbol{U_i} = (u_{i,x}, u_{i,y}, u_{i,z})^T \in R^3$ on the generic model and the corresponding tracked points $\boldsymbol{V_i} \in R^3$ on each range model, we use the radial basis function (RBF) to adapt the generic model on the range face model. The interpolation function is formulated as:

$$f(\boldsymbol{p}) = \boldsymbol{c_1} + [\boldsymbol{c_2 c_3 c_4}] \times \boldsymbol{p} + \sum_{i=1}^{N} \lambda_i \varphi_i (|\boldsymbol{p} - \boldsymbol{U_i}|) \qquad (1)$$

where $\boldsymbol{p_i}$ is a non-control vertex on the generic model and $\boldsymbol{\varphi_i}$ is the RBF for $\boldsymbol{U_i}$. All coefficients $\boldsymbol{c_k}$(k=1,..,4) are determined by solving the equation: $f(\boldsymbol{U_i}) = \boldsymbol{V_i}, i = 1...N$, where $\sum_{i=1}^{N} \lambda_i = 0$ and $\sum_{i=1}^{N} \boldsymbol{U_i} \lambda_i = (0,0,0)^T$. Then, the non-control vertex $\boldsymbol{p_i}$ is mapped to $f(\boldsymbol{p_i})$. The result of adaptation provides sufficient geometric information for subsequent labeling. Figure 2(a) shows an example of an adapted model.

3.2 Geometric Surface Labeling

3D facial range models can be characterized by eight primitive surface features: convex peak, concave pit, convex cylinder, convex saddle, concave saddle, minimal surface, concave cylinder, and planar [25]. After the tracking model is adapted to the range model, each vertex of the adapted model is labeled as one of the eight primitive features. This surface labeling algorithm is similar to the approach described in [16]. The difference is that eight primitive features rather than twelve features are used for our expression representation because we apply a local coordinate system for feature calculation. Let $p = (x, y, z)$ be a point on a surface S, N_p be the unit normal to S at point p, and X_{uv} be a local parameterization of surface S at p. A polynomial patch

$z\left(x,y\right)=\frac{1}{2}Ax^2+Bxy+\frac{1}{2}Cy^2+Dx^3+Ex^2y+Fxy^2+Gy^3$is used to approximate the local surface around p by using X_u , X_v and N_p as a local orthogonal system. We then obtain the principal curvatures by computing the eigenvalues of the Weingarten matrix: $W=(A,B;B,C)$. After obtaining the curvature values of each vertex, we apply the classification method described in [25] to label each vertex of the adapted model. Thus, each range model is represented by a label map $G=[g_1,g_2,...,g_n]$, composed of all vertices' labels on the facial region. Here, g_i is label types and n is the number of vertices in the facial region of the adapted model.

3.3 Optimal Feature Space Transformation

We now represent each face model by its label map *G*. We use LDA to project *G* to an optimal feature space O_G that is relatively insensitive to different subjects while preserving the discriminative expression information. LDA defines the within-class matrix S_w and the between-class matrix S_b. It transforms a *n*-dimensional feature to an optimized *d*-dimensional feature O_G by $O_G = {D_O}^T \cdot G$, where $d < n$, $D_O = arg\left(max_D|\left(D^TS_bD\right)/\left(D^TS_wD\right)\right)$ and *D*, projection matrix. For our experiments, the discriminative classes are 6 expressions, thus the reduced dimension *d* is 5.

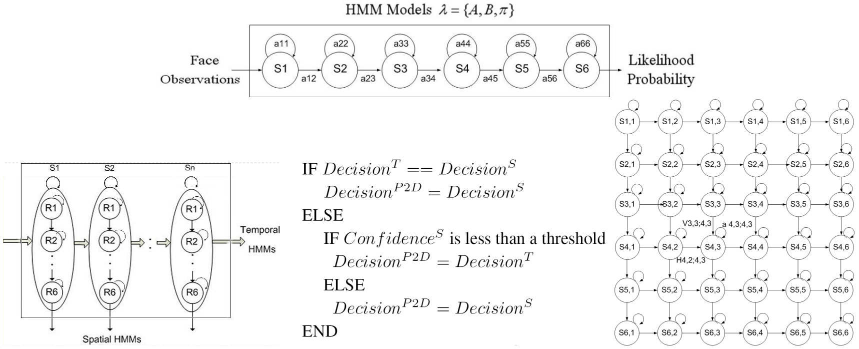
4 HMM Based Classifiers

Facial expression is a spatio-temporal behavior. To better characterize this property, we used Hidden Markov Models to learn the temporal dynamics and the spatial relationships of facial regions. In this section, we describe the Temporal-HMM (T-HMM), Pseudo Spatio-Temporal HMM (P2D-HMM), and real 2D HMM (R2D-HMM), progressively. P2D-HMM is extended from T-HMM, and in turn, R2D HMM is extended from P2D-HMM. As we will discuss, R2D-HMM is the most appropriate method for learning dynamic 3D face models to recognize expressions.

4.1 Temporal HMM

Each prototypic expression is modeled as an HMM. Let $\lambda=[A,B,\pi]$ denote an HMM to be trained and N be the number of hidden states in the model, we denote the states as $S=\{S_1,S_2,...,S_N\}$ and the state at *t* is q_t (see top row of Figure 3). $A=\{a_{ij}\}$is the state transition probability distribution, where $a_{ij}=P\left[q_{t+1}=S_j|q_t=S_i\right]$, $1\le i,j\le N$. $B=\{b_j\left(k\right)\}$is the observation probability distribution in state j, k is an observation . We use Gaussian distributions to estimate each $B=\{b_j\left(k\right)\}$, where $b_j\left(k\right)=P\left[k|q_t=S_j\right]\sim N\left(\mu_j,\Sigma_j\right),1\le j\le N$. Let $\pi=\{\pi_i\}$ be the initial state distribution, where $\pi_i=P\left[q_0=S_i\right],1\le i\le N$. Then, given an observation sequence, $O=O_1O_2...O_T$, where O_i denote an observation at time i, the training procedure is: *Step 1:* Take the optimized feature representation O_G of each observed 3D range face model as an observation. *Step 2:* Initialize the HMM model λ. Each observed model of a sequence corresponds to one state and is used to estimate the parameters in the observation matrix B . Set the initial values of A and π based on observations. *Step 3:* Use the forward-backward algorithm [27] to derive an estimation of the model parameter $\lambda = [A, B, \pi]$ when $P(O|\lambda)$ is maximized. Finally, we derived 6 HMMs; each represents one of the six prototypic expressions.

Given a query model sequence, we follow the Step 1 of the training procedure to represent it as $Q = Q_1 Q_2 ... Q_T$, where the optimized representation of each frame is one observation, denoted as $O_G = (O_{G,1}, O_{G,2}, O_{G,3}, O_{G,4}, O_{G,5})$. Using the forward-backward method, we compute the probability of the observation sequence given a trained HMM_i as $P(Q|\lambda_i)$. We use the Bayesian decision rule to classify the query sequence $c^* = argmax\,[P(\lambda_i|Q)], i \in C$, where $P(\lambda_i|Q) = \frac{P(Q|\lambda_i)P(\lambda_i)}{\sum_{j=1}^{C} P(Q|\lambda_j)P(\lambda_j)}$ and C is the number of the trained HMM models. Since this method trace the temporal dynamics of facial sequences, we denote it as a Temporal HMM(T-HMM). The top row of Figure 3 shows the structure of a 6-state T-HMM. The decision to classify a query sequence to an expression using the T-HMM is denoted as $Decision^T$.



```
IF $Decision^T == Decision^S$
  $Decision^{P2D} = Decision^S$
ELSE
  IF $Confidence^S$ is less than a threshold
    $Decision^{P2D} = Decision^T$
  ELSE
    $Decision^{P2D} = Decision^S$
END
```

Fig. 3. Top:T-HMM; Bottom-left and middle:P2D-HMM and its decision rule; Bottom-right:R2D-HMM

4.2 Pseudo 2D Spatio-temporal HMM

Facial characteristics are not only represented by temporal dynamics (inter-frame) but also by spatial relationships (intra-frame). To model these properties of 3D faces, we investigated the structure of HMMs in the spatial domain combined with the temporal domain, a structure called P2D-HMM.

Spatial HMM (S-HMM): Based on the feature points tracked on the facial surface (e.g., contours of eyebrows, eyes, nose, mouth, and chin), we subdivide each 3D frame model of a sequence into six regions, as shown in Figure 2(b) ($R_1, R_2, ..., R_6$). We then build a 6-state 1D HMM, corresponding the six regions, as shown in each column of P2D-HMM in Figure 3. Similar to the case of entire face regions in the previous section, we transform the labeled map of each sub-region of a frame to an optimized feature space using LDA, denoted as $O_{Gi} = (O_{Gi,1}, O_{Gi,2}, O_{Gi,3}, O_{Gi,4}, O_{Gi,5}), (i = 1..6)$, where *i* is the region index of a frame model. We trained one HMM for each expression. Given a query face sequence with a length N, we compute the likelihood score of each frame and use the Bayesian decision rule to decide the frame's expression type. We make a final decision $Decision^S$ using majority voting. Thus, the query model sequence is recognized as an expression if this expression is the majority result among N frames. As this method tracks spatial dynamics of a facial surface, we call it a spatial HMM (S-HMM).

Combination of Spatial and Temporal HMMs: To model both spatial and temporal information of 3D face sequences, we combine the S-HMM and the T-HMM to construct a pseudo 2D HMM (P2D-HMM) (see Figure 3). The final decision $Decision^{P2D}$ is based on both $Decision^{S}$ and $Decision^{T}$. The decision rule of the P2D-HMM is also described Figure 3. Here, we define $Confidence^{S}$ as the ratio of the number of majority votes versus the total number of frames in the query model sequence. In our experiment, we took 6 frames as a sequence, and chose the threshold for this ratio as 0.67. As a consequence, if at least 4 frames of a query sequence are recognized as expression A by the S-HMM, we determine the query sequence is A. Otherwise, the result comes from the $Decision^{T}$. Essentially, P2D-HMM uses the learned facial temporal characteristics to compensate for the learned facial spatial characteristics.

4.3 Real 2D Spatio-temporal HMM

The aforementioned HMM-based approaches are essentially 1-D or pseudo-2D approaches. However, the dynamic 3D facial models are four dimensional (i.e., 3D plus time). Considering the complexity of high-dimensional HMMs and motivated by the work of Othman *et al* [28] for 2D face recognition, we developed a real 2D HMM (R2D-HMM) architecture to learn the 3D facial dynamics over time. As shown in Figure 3 (bottom-right), this architecture allows for both spatial (vertical) and temporal (horizontal) transition to each state simultaneously. The number of states along spatial (vertical) or temporal (horizontal) axes are all six. Simply put, each 3D sequence contains 6 temporal states, and each frame contains 6 spatial states from top to bottom. The transition from region R_i of the previous frames to another region R_j of the current frame can be learned from the R2D-HMM. In Figure 3, $H_{4,2;4,3}$ and $V_{3,3;4,3}$ are the horizontal and vertical transition probabilities from the state $S_{4,2}$ and the state $S_{3,3}$ to the current state $S_{4,3}$ respectively, and $a_{4,3;4,3}$ is the self-transition probability of the state $S_{4,3}$. Let $O_{r,s}$ be the observation vector of the r^{th} region of the s^{th} frame in a 3D video sequence, the corresponding set of feature vectors is defined as $O_{\{m,n\}} = \{O_{r,s} : 1 \le r \le m, 1 \le s \le n\}$. The feature vector set of the past observation blocks $O_{<m,n>}$is derived by excluding the current observation block $O_{m,n}$, where $O_{<m,n>} = O_{\{m,n\}} - O_{m,n}$. Note that the joint probability of the current state and the observations up to the current observation $P\left(q_{m,n} = S_{a,b}, O_{\{m,n\}}\right)$ can be predicted based on past observation blocks in a recursive form:

$$
\begin{aligned}
&P\left(q_{m,n} = S_{a,b}, O_{\{m,n\}}\right) = P\left(O_{m,n}|q_{m,n} = S_{a,b}\right) \\
&\cdot | \sum_{i,j=1,1}^{M,N} P\left(q_{m,n} = S_{a,b}|q_{m-1,n} = S_{i,j}\right) P\left(q_{m-1,n} = S_{i,j}, O_{\{m-1,n\}}\right) \\
&\cdot \sum_{k,l=1,1}^{M,N} P\left(q_{m,n} = S_{a,b}|q_{m,n-1} = S_{k,l}\right) P\left(q_{m,n-1} = S_{k,l}, O_{\{m,n-1\}}\right) |^{1/2}
\end{aligned}
\qquad (2)
$$

Similar to the standard 1-D HMM, approach, the state matrix is denoted as $\delta_{m,n}\left(a,b\right) = \max_{q_{m-1,m},q_{m,n-1}} P\left[q_{m,n} = S_{a,b}, O_{1,1}, ...O_{m,n}|\lambda\right]$. The observation probability distribution $B_{a,b}\left(O_{m,n}\right)$ is given by

$$B_{a,b}\left(O_{m,n}\right) = \frac{1}{\left[2\pi\right]^{v/2} \Sigma^{1/2}} \cdot e^{\frac{\left(O_{m,n}-\mu_{a,b}\right)\Sigma_{a,b}^{-1}\left(O_{m,n}-\mu_{a,b}\right)^{T}}{2}} \tag{3}$$

Using the Viterbi algorithm, we estimate the model parameter λ as $P\left(O, Q^{*}|\lambda\right)$ is maximized, where $P\left(O, Q^{*}|\lambda\right) = \max_{a,b}\left[\delta_{M,N}\left(a,b\right)\right]$, and Q^{*} is the optimal state sequence. This structure assumes the state transitions to be left-to-right horizontally and top-to-bottom structure vertically. We set the transition matrix in the diagonal direction to be zeros using the same calculation as described in [28]. The expected complexity of the R2D-HMM method is only two times that of the 1D T-HMM structure with the same number of states. In our experiment, given a six-frame sequence, the observation vector is defined by a 6 × 6 matrix *O*, in which each cell is an observation block denoted as $O_{r,s} = \left(O_{r,s,1}, O_{r,s,2}, O_{r,s,3}, O_{r,s,4}, O_{r,s,5}\right)\left(r, s = 1...6\right)$, where *s* is the frame index, *r* is the region index of the frame *s*, and $O_{r,s}(r, s = 1...6)$ is the optimized feature after the label map of the region *r* of the frame *s* is transformed using LDA.

5 Experiments and Analysis

We conducted person-independent experiments on 60 subjects selected from our database. To construct the training set and the testing set, we generated a set of 6-frame subsequences from each expression sequence. To do so, for each expression sequence of a subject, we chose the first six frames as the first subsequence. Then, we chose 6-consecutive frames starting from the second frame as the second subsequence. The process is repeated by shifting the starting index of the sequence every one frame till the end of the sequence. The rationale for this shifting is that a subject could come to the recognition system at any time, thus the recognition process could start from any frame. As a result, 30780 $(= 95 \times 6 \times 54)$ subsequences of 54 subjects were derived for training, and 3420 $(= 95 \times 6 \times 6)$ subsequences of the other 6 subjects were derived for testing. Following a ten-fold cross-validation, we report the average recognition rates of the ten trials as the final result. Our database contains not only the 3D dynamic model sequences but also the associated 2D texture videos. This allows us to compare the results using both 3D data and 2D data of same subjects simultaneously. In the following section, we report the results of our proposed approaches using the 3D dynamic data and their comparative results of the existing approaches using 2D data and 3D static data. All the experiments were conducted in a person-independent fashion.

5.1 Comparison Experiments

Dynamic 3D region-based approaches: We conducted experiments using the Temporal 1D-HMM (T-HMM), Pseudo-2D HMM (P2D-HMM), and Real 2D HMM (R2D-HMM) based on the 3D dynamic surface descriptor. As previously discussed, our facial feature descriptor is constructed from vertices' labels of either entire face region or local facial regions, and we dubbed these methods as "3D *region-based*" approaches. The experimental results are reported in the bottom three rows of the right of Table 1.

Static 2D/3D region-based approaches: *(1) 2D static texture baseline:* We used the Gabor-wavelet based approach [14] as a 2D static baseline method. We used 40 Gabor kernels including 5 scales and 8 orientations and applied them to the 83 key points on the 2D texture frames of all video sequences. *(2) 3D static models baseline:* The LLE based [29], PCA-based, and LDA-based approaches [30] were implemented as the 3D static baseline methods for comparison. The input vector for these three approaches is feature G as explained in section 3.2. For the LLE-based method, we first transform the label map G of each range model to the LLE space and select key frames using k-means clustering. Then, all selected key frame models are used as the gallery models for classification. We use majority voting to classify each 3D query model in the test set. The PCA-based approach and LDA-based approach take the labeled feature G as input vector and apply the PCA and LDA for the recognition. *(3) 3D static models using surface histograms*: We implemented the algorithm reported in [16] as the 3D static baseline method for comparison. We treat each frame of the 3D model sequences as a 3D static model. Based on [16], a so-called primitive surface feature distribution (PSFD) face descriptor is implemented and applied for six-expression classification using LDA. As seen from Table 1, our dynamic 3D model-based HMM approaches outperforms the above static 2D/3D-based approaches. The performance of the PSFD approach is relatively low when it is tested on our 3D dynamic database because its feature representation is based on the static model's surface feature distribution (i.e., histogram). Such a representation may not detect local surface changes in the presence of low-intensity expressions.

Dynamic 2D/3D MU-based approaches: To verify the usefulness of 3D motion units (MU) derived from our dynamic 3D facial models, and to compare it with the 2D MU-based approaches and our dynamic 3D region-based approaches, we implemented the approach reported by Cohen *et al* [9] as the MU-2D baseline method.

(1) *MU-2D based*: According to [9], 12 motion units (MUs) are defined (as the 12 motion vectors) in areas of eyebrows, eye lids, lips, mouth corner and cheeks (see the left three images of Figure 4). Since we have tracked 83 key points on both 2D videos and 3D models as described in Section 3.1, the 12 MU points can be obtained from the tracking result. Note that although more MU points can be used from the tracking (as studied by Pantic *et al* in [8]), to be a fair comparison to the baseline approach, we only used the same 12 MU points as the ones defined in [9]. To compensate for the global rigid motion, we align current frame with the first frame using the estimated head orientation and movement from our adapted 3D tracking model. As such, the displacement vector of a MU point in frame *i* is obtained by $Disp\,(i) = F_i - F_{ne}$, where F_{ne} is the position of the MU point in the first frame (with a neutral expression) and F_i is the position of the MU point in the frame *i*. Figure 4 (left three images) is an example of the 2D MUs derived from a video sequence. In our experiment, we used the 12 MUs, derived from the 2D videos, as the input to the baseline HMM [9] to classify the six prototypic expressions. (2) *MU-3D based*: This is an extension of MU-2D

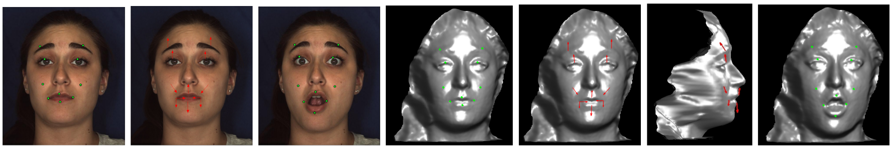


Fig. 4. An example of MUs. Left three: 2D-MUs on the initial frame, motion vectors of MUs from the initial frame to the current frame, and MUs on the current frame of a 2D sequence. Right four: 3D-MUs on the initial frame, 3D motion vectors of MUs with two different views, and MUs on the current frame of a 3D sequence.

method. It derives 3D displacement vectors of the 12 MUs from the dynamic 3D facial videos. Similarly, the 3D model of the current frame is also aligned to the 3D model of the first frame. The compensated 3D motion vectors are then used for HMM classification. Note that although the motion vectors of 2D and 3D models look alike in frontal view, they are actually different since 3D MUs also have motions perpendicular to the frontal view plane, as illustrated in the 2^{nd} image from right of Figure 4. From Table 1, the MU-2D approach achieves a comparable result to that reported in [9] in the case of person-independent recognition. The MU-3D approach outperforms the MU-2D approach because 3D models provides more motion information for FER. Nevertheless, it is not superior to our 3D label-based spatio-temporal approaches because the MU-based approaches do not take advantage of entire facial surface features and rely on very few feature points for classification, and thus are relatively sensitive to the influence of the inaccurate feature detection. The experiment also shows that our 3D label-based R2D-HMM method achieves the best recognition result (90.44%). However, the confusion matrix (Table 2) shows that *sad, disgust*, and *fear* expressions are likely to be mis-classified as *anger*. Our R2D-HMM based approach does not rely on a few features. On the contrary, it takes advantage of entire 3D facial features as well as their 3D dynamics, and thus is more closely matched to the 3D dynamic data and more tolerant to individual feature errors than other compared approaches are.

Table 1. Facial expression recognition results summary

Model property	Method	Recognition rate	Model property	Method	Recognition rate
static 2D	Gabor-wavelet based	63.72%	dynamic 2D	MU-2D	66.95%
static 3D	LLE-based method	61.11%	dynamic 3D	MU-3D	70.31%
static 3D	PCA-based method	70.79%	dynamic 3D	T-HMM based	80.04%
static 3D	LDA-based method	77.04%	dynamic 3D	P2D-HMM based	82.19%
static 3D	PSFD method	53.24%	dynamic 3D	R2D-HMM based	90.44%

Table 2. Confusion matrix using R2D-HMM method

In/out	Anger	Disgust	Fear	Smile	Sad	Surprise
Anger	92.44%	3.68%	1.94%	1.32%	0.00%	1.42%
Disgust	8.28%	87.58%	1.27%	1.27%	0.96%	0.64%
Fear	7.45%	3.42%	85.40%	0.62%	0.00%	3.11%
Smile	0.44%	0.22%	0.66%	97.81%	0.00%	0.87%
Sad	13.12%	1.56%	0.63%	4.06%	80.32%	0.31%
Surprise	0.33%	0.00%	0.00%	0.33%	0.00%	99.34%

5.2 Performance Evaluation Using R2D-HMM

To further evaluate our spatio-temporal based approaches for 3D dynamic facial expression recognition, we conducted experiments to test the robustness of our R2D-HMM method with respect to three aspects: partial facial surface occlusion, expression intensity variation, and 3D model resolution variations.

Partial facial surface occlusion: Limited by views used in our current face imaging system, the facial surface may be partially missing due to the pose variation. To test the robustness of our proposed 3D facial descriptor and the dynamic HMM based classifier, we simulated the situation by changing the yaw and pitch angles of the facial models and generating a set of partially visible surfaces under different views. Ideally, we shall use the ground-true data collected systematically from a variety of views. However, it is hard (as well as expensive) to have such collection due to the difficulty to control the exact degree of pose during the subjects' motion. As such, in this paper we adopt the simulation approach for this study. Such a simulation allows us to study the performance of our proposed expression descriptor in the condition of partial surface invisible with a controllable degree of rotation. For the set of visible surfaces at different orientations, we report the recognition rate separately. Figure 5 shows the facial expression recognition rates with different yaw and pitch angles. The recognition results are based on our proposed dynamic-3D R2D-HMM based approach and the static-3D LDA-based approach. Generally, it shows that our dynamic approach outperforms the static approach in any situation since the motion information compensates for the loss of spatial information.

As shown in the the bottom row of Figure 5, our approach achieves a relatively high recognition rate (over 80%) even when the yaw and pitch angles change to 60 degrees, which demonstrates its robustness to the data loss due to the partial data invisible. The first row of Figure 5 shows the FER rate when the pose changes in only one dimension (yaw/pitch). Out of the useful range (i.e., either pitch or yaw angle changes exceed 150 degrees from the frontal view), the FER rate degrades to zero dramatically because of the paucity of useful information for recognition. The recognition curve of yaw's rotation within the useful range (Top-Left of Figure 5) is approximately symmetric with respect to the zero yaw angle. The recognition rate does not decrease too much even when the yaw angle is close to 90-degree (corresponds to half face visible). This is because either the left part or the right part of a face compensates for the other in the 3D space due to the approximate symmetric appearance of the face along the nose profile. However, the recognition curve of tilts rotation within the useful range is a little asymmetric as shown in the Top-Right of Figure 5. When the face is tilted up, the recognition rate is degraded not as much as when the face is tilted down in the same degree. This asymmetric property implies that the lower part of the face may provide more useful information than the upper part for expression recognition.

Variation of expression intensity: Our approach can also deal with variations of expression intensity since it not only includes different levels of intensities but also considers their dynamic changes. Based on our observation, we simply separated each 3D video sequence into two parts: a low intensity sequence (e.g., subsequences close to the starting or ending frames showing near-neutral expressions) and a high intensity

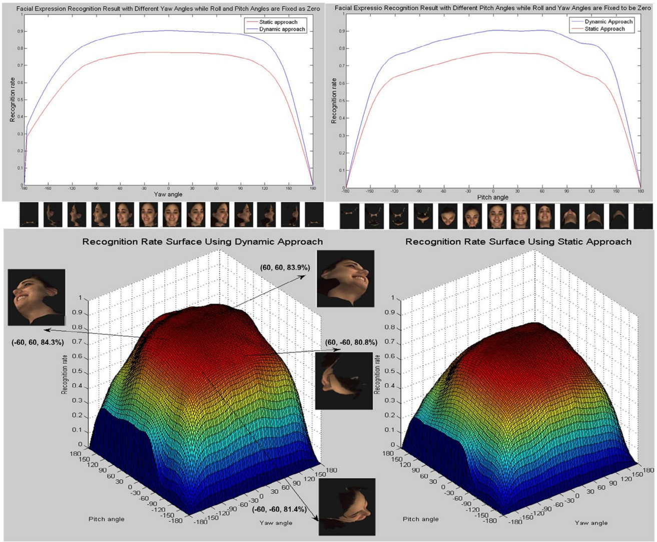


Fig. 5. FER results with simulated partial data missing scenario. Top: FER rate curves with respect to yaw rotation only and pitch rotation only; Bottom: FER rates surface with both yaw and pitch rotations. The facial pictures in the bottom illustrate the visible parts of a face when the yaw and pitch angles change to +/- 60 degrees. The recognition rates are also denoted besides the pictures.

sequence (subsequences excluding the low-intensity sequences). We performed the test on the low-intensity and high-intensity expressions individually using the R2D-HMM approach and the static PSFD approach [16]. Our training set includes both levels of intensities. The results show that the R2D-HMM method can detect both weak and strong expressions well. It achieves a 88.26% recognition rate of low intensity expressions and 91.58% recognition rate of high intensity expressions. However, the PSFD method has 71.72% recognition rate of high intensity expressions. It has less than 50% recognition rate for low intensity expressions. The main reason is that the static surface histogram descriptor may not be able to capture small variations of facial features as our 3D surface label descriptor does. In addition, the high performance of our approach is also attributed to the applied R2D-HMM classifier, which learns temporal transitions of dynamic facial surfaces effectively for both low-intensity and high-intensity expressions.

Variation of facial model resolutions: We down-sampled the test models to a low-resolution version with around 18,000 vertices, which is almost half the resolution of the original facial models (35,000 vertices) used for training. We then conducted the experiment to see whether the proposed approach works well for facial models with different resolutions. Based our R2D-HMM approach, the recognition rate for the low resolution models is 89.78%, which is comparable to the result of high resolution models (90.44%). This demonstrates that our approach has certain robustness to different resolutions, despite the fact that different resolutions could blur or sharpen the shape of facial surface. This result is supported by the psychological finding: blurring the shape information has little effect on the recognition performance as long as the motion channel is presented [31].

6 Discussions and Conclusions

In this paper, we proposed a spatio-temporal approach to study the viability of using dynamic 3D facial range models for facial expression recognition. Integrating the 3D facial surface descriptor and the HMMs (R2D-HMM, or P2D-HMM, or T-HMM), our system is able to learn the dynamic 3D facial surface information and achieves 90.44% person-independent recognition rate with both low and high intensities. In general, the HMM has been widely used for 2D facial expression recognition and face recognition. However, the way that we applied the real 2D-HMM to address 3D dynamic facial expression recognition is novel. We have extended the work of FER from static 3D range data to 3D videos. Many previous studies showed that sequential images are better than static images for FER [9,7]. We have verified that this statement holds true for 3D geometric models. The advantage of our 3D dynamic model based approach has been demonstrated as compared to several existing 2D static/video based and 3D static model based approaches using our new 3D dynamic facial expression database. This database will be made public to the research community. Ultimately, however, our focus was to study the usefulness of the new dynamic 3D facial range models for facial expression recognition rather than develop a fully automatic FER system. Our current work requires a semi-automatic process to select feature points at the initial stage. A fully automatic system with a robust 3D feature tracking will be our next stage of the development. To investigate the recognition performance in terms of large pose variations, we will design a new approach to measure the exact pose degree during the capture of ground-true spontaneous expressions. In addition, we will also investigate an approach to detect 3D action units and integrate the motion vector information with our surface label descriptor in order to improve the current FER performance.

Acknowledgement

This material is based upon the work supported in part by the National Science Foundation under grants IIS-0541044, IIS-0414029, and the NYSTAR's James D. Watson Investigator Program.

References

1. Ekman, P., Friesen, W.: The Facial Action Coding System. Consulting Psychologists Press, San Francisco (1978)
2. Tong, Y., Liao, W., Ji, Q.: Facial action unit recognition by exploiting their dynamic and semantic relationships. IEEE Trans. on PAMI 10, 1683–1699 (2007)
3. Pantic, M., Rothkrantz, L.: Automatic analysis of facial expressions: the state of the art. IEEE Trans. PAMI (2000)
4. Yang, P., Liu, Q., Metaxas, D.: Boosting coded dynamic features for facial action units and facial expression recognition. In: CVPR 2007 (2007)
5. Bartlett, M., et al.: Fully automatic facial action recognition in spontaneous behavior. In: FGR 2006, pp. 223–228 (2006)

6. Donato, G., Bartlett, M., Hager, J., Ekman, P., Sejnowski, T.: Classifying facial actions. IEEE Trans. PAMI 21(10), 974–989 (1999)
7. Lien, J., et al.: Subtly different facial expression recognition and expression intensity estimation. In: CVPR 1998 (1998)
8. Pantic, M., Patras, I.: Detecting facial actions and their temporal segments in nearly frontal-view face image sequences. In: IEEE Int'l Conf. on Systems, Man and Cybernetics 2005, pp. 3358–3363 (2005)
9. Cohen, I., Sebe, N., Garg, A., Chen, L., Huang, T.: Facial expression recognition from video sequences: temporal and static modeling. Journal of CVIU 91 (2003)
10. Sebe, N., et al.: Authentic facial expression analysis. Image Vision Computing 12, 1856–1863 (2007)
11. Zeng, Z., et al.: Spontaneous emotional facial expression detection. Journal of Multimedia 5, 1–8 (2006)
12. Chang, Y., Hu, C., Turk, M.: Probabilistic expression analysis on manifolds. In: IEEE Inter. Conf. on CVPR 2004 (2004)
13. Zhao, G., Pietikainen, M.: Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE Trans. on PAMI (2007)
14. Lyons, M., et al.: Automatic classification of single facial images. IEEE Trans. PAMI 21, 1357–1362 (1999)
15. Zalewski, L., Gong, S.: Synthesis and recognition of facial expressions in virtual 3d views. In: FGR 2004 (2004)
16. Wang, J., Yin, L., Wei, X., Sun, Y.: 3d facial expression recognition based on prmitive surface feature distribution. In: IEEE CVPR (2006)
17. Wang, P., Verma, R., et al.: Quantifying facial expression abnormality in schizophrenia by combining 2d and 3d features. In: CVPR 2007 (2007)
18. Wang, Y., Huang, X., Lee, C., et al.: High resolution acquisition, learning and transfer of dynamic 3d facial expressions. In: EUROGRAPHICS 2004 (2004)
19. Di3D, I. (2006), http://www.di3d.com
20. Chang, Y., Vieira, M., Turk, M., Velho, L.: Automatic 3d facial expression analysis in videos. In: IEEE ICCV 2005 Workshop on Analysis and Modeling of Faces and Gestures (2005)
21. Yin, L., Wei, X., Sun, Y., Wang, J., Rosato, M.: A 3d facial expression database for facial behavior research. In: IEEE FGR 2006 (2006)
22. Yeasin, M., et al.: From facial expression to level of interest: a spatio-temporal approach. In: CVPR 2004 (2004)
23. Phillips, P., Flynn, P., et al.: Overview of the face recognition grand challenge. In: CVPR 2005 (2005)
24. Yin, L., Chen, X., Sun, Y., Worm, T., Reale, M.: A high resolution 3d dynamic facial expression database. In: IEEE FGR 2008 (2008)
25. Sun, Y., Yin, L.: 3d face recognition using two views face modeling and labeling. In: IEEE CVPR 2005 Workshop on A3DISS (2005)
26. Cootes, T., Edwards, G., Taylor, C.: Active appearance models. IEEE Trans. PAMI 23 (2001)
27. Rabiner, L.: A tutorial on hidden markov models and selected applications in speech recognition. Proceedings of IEEE, 77(2) (1989)
28. Othman, H., Aoulnasr, T.: A separable low complex 2d hmm with application to face recognition. IEEE PAMI 25 (2003)
29. Saul, L., Roweis, S.: Think globally, fit locally: Unsupervised learning of low dimensional manifolds. Journal of Machine Learning Research 4, 119–155 (2003)
30. Martinez, A., Kak, A.: Pca versus lda. IEEE Trans. on PAMI 23, 228–233 (2003)
31. Wallraven, C., et al.: Psychophysical evaluation of animated facial expressions. In: Proc. of the 2nd Symposium on Applied Perception in Graphics and Visualization (2005)