Feature Detection and Tracking on Geometric Mesh Data Using a Combined Global and Local Shape Model for Face Analysis

Shaun Canavan and Lijun Yin

State University of New York at Binghamton

Abstract

Automatic geometric feature localization is the first step towards the 3D based face analysis. In this paper we propose a shape model with a local and global constraint for feature detection. Such a so-called shape-index based statistical shape model (SI-SSM) makes use of the global shape of the facial data as well as local patches, consisting of shape index values, around landmark features. The fitting process and the performance of our proposed method are evaluated in terms of various imaging conditions and data qualities. The efficacy of the detected landmarks is validated through applications for geometric based face identification.

1. Introduction

Geometric feature localization on 3D/4D range data is crucial for geometric based face modeling and recognition [21][22][28][2]. While 2D based tracking methods have been successfully developed, such as Active Appearance Models [4], and Constrained Local Models (CLM) [5], there is a need for novel and robust algorithms to handle 3D/4D range data. There has been recent work to address the problem of detecting feature landmarks on range data. Zhao et al. [20] had success with detecting 3D landmarks using a statistical facial feature model; however there is an upper bound on the number of landmarks. Fanelli et al. [7] used an active appearance model that is based on random forests; however this method used depth and intensity data rather than the 3D/4D range data. Nair et al. [11] fit a 3D active shape model to facial data using candidate landmarks to deform the model, however the resulting error rate for fitting is relatively large, and problems occur when holes exist around the nose Perakis et al. [12] used a 3D active shape model which was fit from previously determined candidate landmarks. A drawback to this method is the need for preprocessing. Jeni et al. [9] used a 3D constrained local model method (estimated from 2D shape) to track landmarks for action unit intensity estimation. The work by Sun et al. [15] tracked features in the 2D space and the 3D features themselves were obtained by mapping the 2D features to the corresponding parts of the 3D models. Our previous work [3] using a 3D temporal deformable shape model shows the limitation in moderated facial motion.

Motivated by the work [11] [3] and the shape descriptor [6], we propose to model the statistical shape on a shape-index domain with both global and local constraints. The so-called *shape index-based statistical shape model (**SI-SSM**)* is constructed from both the global shape of 3D feature landmarks and local features from patches around each landmark. In order to construct the patches we search 3D features from the (*u, v*) coordinates around each landmark. From these new features we construct a $n \times n$ patch, where each vertex is represented by a unique shape index value. Using both the global shape and the local features around each landmark enables us to reliably detect and track features on the range mesh data. The feature detection and tracking are based on the correlation between the local shape index patches on the input range data and the trained SI-SSM model. To validate the usefulness of the feature detection method, we present a novel 3D facial descriptor, a so-called spatio-temporal shape-index feature descriptor (ST-SIFD), for application of face classification.

This work contributes to applying a statistical model that makes use of both the global shape of 3D face surface, as well as the local shape around individual features by way of shape index representation. The local fitting with shape index representation allows for feature detection under various pose and illumination changes. We are able to model and fit data that include various rotations, expressions, occlusions, and missing data by training on each of these data types. Moreover, the proposed ST-SIFD facial representation characterizes individual identities for recognition purpose.

The paper is organized as follows. Section 2 will present the shape index-based statistical model with local and global features combined. Section 3 will describe the feature detection and tracking algorithm. Experiments and performance evaluation will be given in Section 4, followed by a case study for face identification application.

2. SI-SSM

Our proposed method models both the global shape of

3D facial landmarks, as well as the local curvatures from patches around the landmarks. In order to construct the SI-SSM, we annotate the training data with L landmarks. From these annotated landmarks we are able to model both the global and local shapes of a face. The resulting global shape, local curvature patches, and the final construction of the SI-SSM are detailed in the following sub-sections.

2.1. Global face shape

To model the global face shape, we first create a $n \times n$ patch around each of the L annotated landmarks for each training mesh. To construct these patches we use the corresponding (u, v) coordinates for each of the training data. Given a set of M training data, each with L patches, a parameterized model, S_G, is constructed, where $S_G = (x_1, y_1, z_1, \dots, x_N, y_N, z_N)$, $N = L \times n \times n$. This parameterized model contains the global shape of all of the training data. The first step to create this model is aligning the N landmarks, on each of the M training data, by using a modified version of Procrustes analysis [4]. PCA is then applied to learn the modes of variation from the training data. We can then approximate any shape by

$$S_G = \bar{s} + Vw \quad (1)$$

where $\bar{s}$ is the mean shape, V is the eigenvectors of the covariance matrix C, which describes the modes of variation learned from the training data, and w is a weight vector used to generate new shapes (SI-SSM instance) by varying its parameters within certain limits. For our model, we constrain those valid shapes in the range between two standard deviations from the mean.

A smaller constraint would shrink the search space and possibly miss the best fit to the input model. A larger domain would create an unnecessarily large search space that would have instances of the model that do not look like a face.

2.2. Local face shape

To model the local face shape we apply the shape index values to represent the local patches. To do so, we calculate the shape index values for each of the L patches in the global face shape. Calculating the shape index [6] gives us a quantitative measure of the shape of each patch around the L annotated landmarks.

We normalize the shape index scale to [0, 1] and encode them as a continuous range of grey-level values between 1 and 255. To give us an efficient description of the data, we transform the shape index scale to a set of nine quantization values from concave to convex.

Given the set of M training data with L patches where each contains the calculated shape index values, we construct a second parameterized model $S_L = (SI_1, \dots, SI_N)$. PCA is then applied to this local shape vector in the same manner as the global shape vector does. We construct a new vector, V_{SI}, which yields of the modes of variation along the principal axes for the local shape index values. Similar to the global shape, we can approximate any local patch shape using the vector, V_{SI}, and a weight vector w_{SI} by

$$S_L = \bar{si} + V_{SI} w_{SI} \quad (2)$$

2.3. Combined global and local feature model

In order to integrate the two features into a combined feature model, we concatenate both the global and local shape feature vectors into one feature vector $\boldsymbol{S_{GL}}$, where $\boldsymbol{S_{GL}} = (x_1, y_1, z_1, \dots, x_N, y_N, z_N, SI_1, \dots, SI_N)$. Such a combination allows us to move the local patches to a more representative surface on the model while maintaining the constraint in the allowable shape domain. Other methods that use statistical models such as [4][5][15] have been successful in using statistical models to create a combined feature vector using both the shape and "appearance" of the face. The "appearance" portion (e.g. textures) of the model helps to guide the model and fit to new data, however, these approaches suffer from the problem of global lighting variation, as well as skin tone of the modeled face. The grey-level appearance information in these models must be normalized in order to handle this lighting variation. Our SI-SSM uses shape index values to model the local features, which guide our model and fit to new range data. Shape index values are relatively invariant to global lighting variation and skin tone. It is a quantitative measure of shape, by using these features our model does not encounter the same issues that similar "appearance" based solutions do.

3. Landmark Detection and Tracking

In order to perform the detection and tracking of landmarks, we must first calculate the shape index values for the vertices of the input range mesh. This is done in the same manner as described in Section 2.2. The SI-SSM fitting algorithm is described below.

First, an initialization phase is performed to give us a sufficient starting point to perform a local patch-based correlation search. During the initialization phase we learn the weight parameters w of the global shape by uniformly varying the weight vector to generate new instances of the SI-SSM. Iterative Closest Point (ICP) [1] is used to minimize the distance between each SI-SSM instance and the input range data. The patches from the instance of the SI-SSM with the lowest ICP matching score are used as the initialized starting landmarks for the SI-SSM. Given this global fit, we then calculate the local patch–based correlation score between the SI-SSM and the input range

mesh, which is computed using a cross correlation template matching scheme [10].

The final correlation score, *CS*, is computed as

$$CS = \sum_{p=1}^{L} CS_p \quad (3)$$

which allows us to have a base line comparison for the local patch-based correlation search, as well as define a tighter convergence criterion.

Once we have the initialized patches and initial correlation score we then perform a local search around each of the patches of the SI-SSM. For each patch in our model we construct a new patch of the same size around each of the $n \times n$ points of the original patch. For example, when $n=3$, we construct a patch centered on each point of the original 3×3 patch, resulting in 9 new patches. The shape index values for each of these patches correspond to the shape index values of the vertices of the new patches. We compute a new CS_p for each of the new patches we created. The patch with the highest correlation score is labeled as the new SI-SSM patch. It is important to make sure that when all of the patches have been moved the new global shape of the face is with the allowable shape domain of ± 2 standard deviations from the mean. From formula 3, we can derive the corresponding w_{SI} vector of the newly transformed SI-SSM by the following:

$$w_{SI} = V_{SI}^{T}(S_L - \bar{sl}) \quad (3)$$

This new weight vector is constrained to be within the allowable shape domain, and we approximate a new shape by again utilizing formula 3 with this weight vector.

Once we have the new approximated global shape of the face, iterative closest point is then used to again minimize the distance between the new SI-SSM instance and the range mesh. This process continues until convergence is reached. Notice that the convergence is defined by two main criteria:

1) The computed correlation score, CS, for the transformed SI-SSM is higher than the score in the previous iteration.
2) The computed correlation score, CS, exhibits little to no change from the CS computed in the previous iteration.

Once we have the detected features for the current 3D mesh in the sequence, we then use the iterative closest point algorithm to move the landmarks to the next mesh in the sequence and continue the tracking of the sequence. The fitting process is then repeated with the previously detected landmarks used as the initial model fit. Figure 1 shows several sample 3D dynamic models with detected feature vertices.

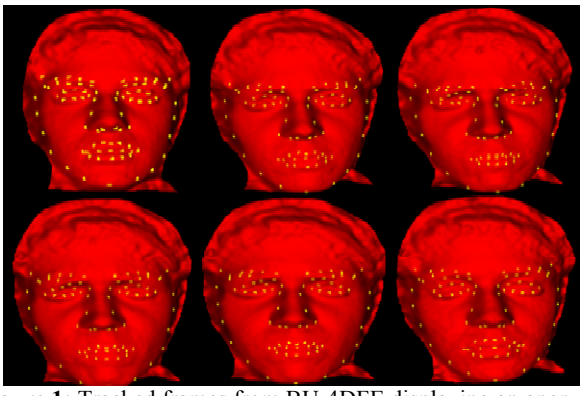

Figure 1: Tracked frames from BU-4DFE displaying an angry expression.

4. Experiments and Evaluation

We have tested out algorithm on public BU-4DFE [18] and BP4D-Spontaenous [19] datasets. Those datasets include 101 subjects with posed expressions for BU-4DFE and 41 subjects with spontaneous expressions for BP4D.

4.1. Accuracy of Feature Detection

To evaluate the accuracy of detecting and tracking landmarks using our SI-SSM method, we calculate the mean squared error between the ground truth and our detected/tracked landmarks (centroids of patches). We do this by calculating the one-point spacing between each landmark. The one-point spacing is defined as the closest pair of points on the 3D scans (0.5mm on the geometric surface). We treat the unit error as equal to 1 point spacing, so we can compute the average of the point distances between the sets. The results show that the error rate, mean squared error (MSE), is 3.2 for 4DFE and 2.9 for BP4D-Spontaneous. Note that the ground truth feature points that have been used for comparison are provided by 4DFE and BP4D-Spontaneous databases ($L=83$).

We also compared our SI-SSM algorithm, using 10% of the data for training and the rest for testing, against other state of the art algorithms. We compared our results against a 2D CLM mapped to 3D [9], TDSM [3], and Sun et al. [15] on the BU-4DFE and BP4D databases. For all comparisons we use the centroid landmark in each of the patches for comparisons, and report the MSE of the average point spacings. Note that the data tracked with the 2D CLM only used 66 landmarks while we used 83 landmarks. In order to perform these experiments we selected the common sub-set of the two sets of landmarks, resulting in 49 landmarks for comparison. These landmarks comprise of the left and right eyes, nose, mouth and landmarks on the contour of the face. The 3D features mapped from 2D CLM have an error rate of 13.2 as compared to ours of 2.9. The high error rate of the 2D CLM based method can be attributed to frames where the tracking was lost and the method was unable to find a

correct fit, as well as the mapping error from 2D to 3D. Figure 4 (Bottom-right) shows an example where the 2D CLM based method was unable to detect the correct landmarks while our SI-SSM was successful in detecting them. As can be seen from Table 1, which shows the results from these comparisons, our SI-SSM method outperforms the compared state of the art methods.

MSE	BU-4DFE	BP4D-Spontaneous
3D Mapped From 2D CLM [9]	N/A	13.2
TDSM [3]	3.7	4.0
Sun [15]	6.3	7.2
SI-SSM	**3.2**	**2.9**

Table 1: Comparison of SI-SSM, TDSM, Sun, and 2D CLM Mapped to 3D.

4.2. Performance Evaluation

We also conducted performance evaluation for feature detection and tracking in terms of head pose rotation and low quality data with incomplete surfaces. The details and errors statistics are detailed in the following sub-sections.

4.2.1 Head pose rotation

We also tested our algorithm on sequences that contain rotations only. Figure 2 shows several examples of rotations between [0, 90]. For all tested rotation sequences there is a MSE of 3.2. Figure 3 shows the average point spacings error along with the standard deviation for each of the, *L=83*, landmarks. As can be seen in Figure 3 (Top) the average error for rotations is fairly stable across all for all landmarks degrees, showing robustness to large rotations. Figure 3 (Bottom) shows the average error, for all landmarks, across each of the degrees in the range [0, 90].

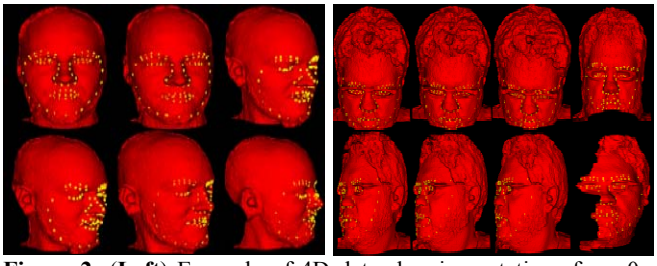

Figure 2: (**Left**) Example of 4D data showing rotations from 0 degree to 90 degree; (**Right**) Sample frames displaying pitch and yaw pose estimations. Upper Row (Pitch): -20, -23, -27; Lower Row (Yaw): -37, -49, -51; *Note: The last column is the same model from the previous column.*

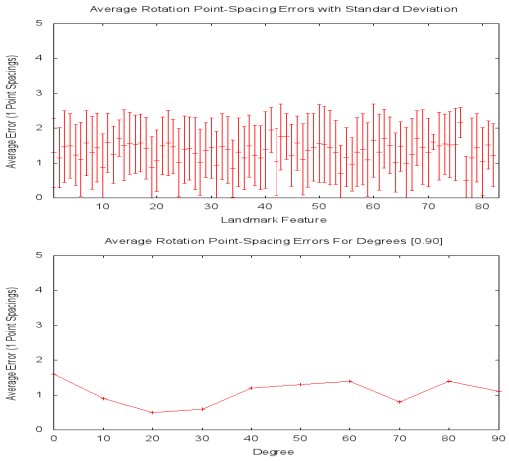

Figure 3: (**Top**) Average error for sequences displaying occlusions from rotations; (**Bottom**) Average error in relation to rotation degree.

The results displayed in Figure 2 demonstrate the SI-SSM is robust to large rotations. For all rotations the average point spacing error is fairly consistent remaining under 2, with the largest error being 1.6 and the smallest error being 0.5.

4.2.2 Low quality data

We also tested the accuracy of our algorithm on sequences that contain low quality data. We define low quality data as missing data from self-occlusion (eye glasses, hand in front of face, etc.), noisy data (beards, distorted patches caused by 3D data capture, etc.), and incomplete scans containing holes and isolated patches. Figure 4 illustrate these types of low quality data sequences.

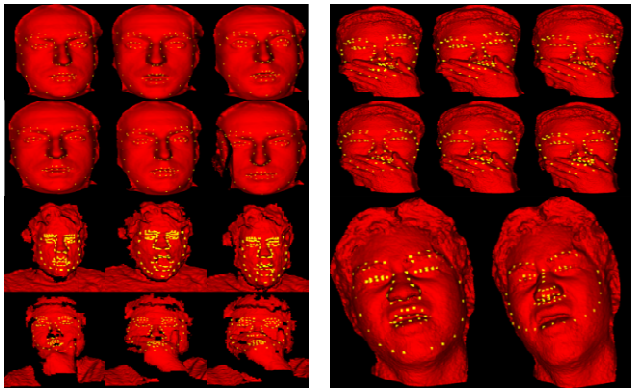

Figure 4: (**Top-left**) Tracked frames from a surprised expression. **NOTE**: the bottom right frame in the sequence is missing data, and the SI-SSM still fits to the missing data; (**Top-right**) Robustness to occlusion; (**Bottom-left**) Robustness to noise and missing data; (**Bottom-right**): Correct fit by SI-SSM (left face), erroneous fit by 3D mapped from 2D CLM (right face). NOTE: In this sequence there are approximately 100 frames that the 2D CLM fails to correctly fit whereas the SI-SSM is successful.

Our test results on low quality data shows the MSE error is 3.6 on average. While the MSE is slightly higher and the average errors show more variance than other tested data, the SI-SSM is still able to successfully fit to this data with a generally low error rate, showing robustness to low quality data.

5. Applications

To validate our proposed method, we apply it to the subject identification problem. We take the component based approach for classification. Given the detected facial landmarks, local regions can be formed in the areas of eyebrow, eyes, nose, mouth, and cheek, etc. (similar to [26]). Therefore, the shape change along the 3D model sequences can be observed in individual local regions. Inspired by the existing work on facial analysis using curvature based approaches [16][26] and dynamic texture based approaches [27][24], we apply a component-based spatio-temporal shape-index feature descriptor (ST-SIFD) for face representation and classification.

We transform the shape index scale to a set of nine quantization values from concave to convex, namely (1) Cup (0); (2) Trough (0.125); (3) Rut Saddle (0.25); (4) Rut (0.375); (5) Saddle (0.5); (6) Saddle Ridge (0.625); (7) Ridge (0.75); (8) Dome (0.875); and (9) Cap (1). Based on the shape index computation [10][6], each vertex on the 3D face model is assigned a quantized SI value. Each facial model is segmented into sub-regions, similar to the regions defined in [26], e.g., eyebrow, eyes, nose, mouth, cheek, etc. From the set of detected feature points, we are able to derive the spatio-temporal SI distribution of each sub-region and combine them into a vector. The ST-SIFD is constructed as follows.

5.1. Facial feature descriptor: ST-SIFD

In the 3D geometric space, each individual region h_i ($i=1,2,..,8$) is observed (and concatenated) across k consecutive frames, constructing a regional 3D volume V_i. Given the total number of vertices N_i in the 3D volume V_i, and the number of vertices m_j with shape-index scale j, ($j=1,...,9$) in that volume, the histogram v_i of each volume is derived as follow:

$$v_i = [\frac{m_1}{N_i}, \frac{m_2}{N_i}, ..., \frac{m_9}{N_i}] \quad (4)$$

As a result, the spatio-temporal feature descriptor g is the concatenation of the histogram v_i of all eight regional volumes:

$$g = [v_1, v_2 ..., v_8] \quad (5)$$

Note that the size of regional 3D volume is determined by the number of consecutive frames (k) which will be concatenated. In our experiment, we take $k=6$ for data with dynamic facial models.

5.2. Classification

To reduce the high dimensionality of the spatio-temporal feature vector g, we apply a linear discriminant analysis (LDA) based method to reduce the feature space. The LDA transformation maps the feature space into an optimal space for face classification. It transforms the n-dimensional feature g to the d-dimensional optimized feature O_g ($d<n$) .

We use the Support Vector Machine (SVM) for classification of subjects. Traditional SVM is used for binary classification. For multi-class situations, the one-against-one SVM method (a.k.a. one-versus-one method) can be used. An SVM is constructed for every pair of classes by training it to discriminate the two classes. A max-min strategy is used to determine the class that a test sample belongs to. That is to say, the class with maximum number of votes for the test sample is assigned to the sample.

Alternatively, one efficient way is to construct a multi-class classifier by combining several binary classifiers. The one-against-all SVM is constructed for each class by discriminating that class against the remaining classes. A test pattern x is classified by using the winner-takes-all decision strategy, i.e., the class with the maximum value of the discriminant function $f(x)$ is the class that x belongs to. Considering the algorithm complexity and classification performance, we chose the one-against-all SVM for the classification experiment.

5.3. Experiment on subject identification

Subsets of face databases 4DFE and BP4D are used for face identification experiment.

(i) Dynamic 3D sequences (4DFE): We chose 30 subjects for experiment. For each subject, 10 sets of six consecutive frames of each expression were randomly chosen, resulting in 60 sample sets with six expressions and 1,800 sample sets for 30 subjects in total for training. The frames $k=6$ are used for constructing the feature vector. A randomly selected 6-frame test set (six for each subject) was used for testing. The recognition rate is 92.7%.

(ii) Dynamic 3D sequences (BP4D-Spontaneous): We also chose 30 subjects for experiment. For each subject, 10 sets of six consecutive frames of each expression were randomly chosen. With eight expressions and 30 subjects, there are 2,400 sample sets in total for training. The frames $k=6$ are used for constructing the feature vector. A randomly selected 6-frame test set (eight for each subject) was used for testing. The recognition rate is 93.4%.

6. Conclusion and Future Work

In this paper we have presented a novel method of detecting and tracking landmarks on 3D and 4D data using a shape index-based statistical shape model. The SI-SSM has been tested on the public 3D dynamic face databases. The SI-SSM has shown feasible of handling cases in various data qualities. It also shows the improved performance as compared to the peer approaches, given only the geometric information used. The utility of the proposed approach has been validated through ST-SIFD based face classification.

In our future work, we will investigate the influence of the patch size on the fitting results. We will also investigate other optimal geometric features with a combination of texture information for applications in real-time face analysis. In addition, the evaluation and validation test will be conducted on more datasets with a much larger scale in the future.

7. Acknowledgement

This material is based upon work supported in part by the National Science Foundation under grants IIS-1051103, CNS-1205664, and IIS-0541044.

References

[1] P. Besl, N. McKay, "A method of registration of 3D shapes," IEEE Trans. on PAMI vol. 14, pp. 239-256, 1992.

[2] V. Blanz, T. Vetter, "A morphable model for the synthesis of human faces," SIGGRAPH, 1999.

[3] S. Canavan, X. Zhang, L. Yin, "Fitting and tracking 3D/4D facial data using a temporal deformable shape model," IEEE International Conference on Multimedia and Expo 2013.

[4] T. Cootes, K. Walker, and C.J. Taylor, "View-based active appearance models," Image and Vision Computing 20(9): 657-664, 2002

[5] D. Cristinacce and T. Cootes, "Feature Detection and Tracking with Constrained Local Models," BMVC, 2(5), 2006.

[6] C. Dorai and A. Jain, "Cosmos – A Representation Scheme for 3D Free-form Objects," IEEE Trans. PAMI, (19)10, 1997.

[7] G. Fanelli, M. Dantone, and L.V. Gool, "Real time 3D face alignment with random forests-based active appearance models," IEEE International Conference on Automatic Face and Gesture Recognition (FGR'13), 2013.

[8] P. Flynn and A. Jain, "Surface classification: Hypothesis testing and parameter estimation," IEEE CVPR 1988.

[9] L. Jeni, A. Lorinca, et al, "3D shape estimation in video sequences provides high precision evaluation of facial expressions, "Image and Vision Computing, (30)10: 785-795, 2012.

[10] J. Lewis, "Fast Template Matching," Vision Interface. 1995.

[11] P. Nair, A. Cavallaro, "3D face detection landmark localization, and registration using a point distribution model," IEEE Trans. on Multimedia (11), 2009.

[12] P. Perakis, G. Passalis, T. Theoharis, and I. Kakadiaris, "3D Facial Landmark Detection Under Large Yaw and Expression Variations," IEEE Trans. on PAMI, 35(7):1552-1564, 2013.

[13] P. Phillips, P. Flynn, T. Scruggs, K. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, J. Worek, "Overview of the face recognition grand challenge," Computer Vision and pattern Recognition 2005.

[14] Rabiner, L.: A tutorial on hidden markov models and selected applications in speech recognition Proceedings of IEEE, 77(2), 1989.

[15] Y. Sun, X. Chen, M. Rosato, and L. Yin, "Tracking vertex flow and model adaption for 3D spatio-temporal face analysis," IEEE Trans. on SMC-A, 40(3):461-474, 2010.

[16] J. Wang, L. Yin, X. Wei, Y. Sun, "3D facial expression recognition based on primitive surface feature distribution," IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), 2006.

[17] L. Yin, X. Wei, Y. Sun, J. Wang, M. Rosato, M., "A 3D facial expression database for facial behavior research," IEEE International Conference on Automatic Face and Gesture Recognition (FGR'06), 2006.

[18] L. Yin, X. Chen, Y. Sun, T. Worm, M. Reale, "A high-resolution 3D dynamic facial expression database," IEEE International Conference on Automatic Face and Gesture Recognition (FGR'08), 2008.

[19] X. Zhang, L. Yin, J. Cohn S. Canavan, M. Reale, A. Horowitz, P. Liu, and J. Girard, "BP4D-Spontaneous: A high resolution 3D dynamic facial expression database," Image and Vision Computing, 32, p692-706, 2014.

[20] X. Zhao et al, "Accurate landmarking of three-dimensional facial data in the presence of facial expressions and occlusions using a three-dimensional statistical facial feature model," IEEE Trans. SMC Part B, 41(5): 1417-1428, 2011.

[21] I. Kakadiaris, G. Passalis, G. Toderick, T. Theohari, et al, "Three-dimensional face rec. in the presence of facial expressions: An annotated deformable model approach," IEEE Trans. PAMI 29, 2007.

[22] M. Segundo, C. Queirolo, O.R.P. Bellon, L.Silva, "Automatic 3D facial segmentation and landmark detection," International Conference on Image Analysis and Processing 2007.

[23] J. Sun, D. Huang, Y. Wang, and L. Chen, A Coarse-to-Fine Approach to Robust 3D Facial Landmarking via Curvature Analysis and Active Normal Model, IEEE International Joint Conference on Biometrics, 2014.

[24] S. Koelstra, M. Pantic, and I. Patras, A dynamic texture-based approach to recognition of facial actions and their temporal models. *IEEE Trans. on PAMI*, 32(11), 2010.

[25] E. Ong, and R. Bowden, "Robust facial feature Tracking using shape-constrained multiresolution-selected linear predictors," IEEE Trans. on PAMI, 33(9):1844-1859, 2011.

[26] S. Canavan, Y. Sun, and L. Yin, "A Dynamic Curvature Based Approach for Facial Activity Analysis in 3D Space", *IEEE CVPR Workshop on SISM 2012*.

[27] G. Zhao and M. Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Trans. PAMI,* 6(29), 2007.

[28] G. Sandback, S. Zafeirious, M. Pantic, and L. Yin, "Static and dynamic 3D facial expression recognition: A comprehensive survey," Image and Vision Computing (30)10: 683-697, 2012.