Magnifying subtle facial motions for 4D Expression Recognition
Qingkai Zhen, Di Huang, Yunhong Wang, Hassen Drira, Ben Boulbaba, Mohamed Daoudi

To cite this version:
Qingkai Zhen, Di Huang, Yunhong Wang, Hassen Drira, Ben Boulbaba, et al.. Magnifying subtle facial motions for 4D Expression Recognition. 23rd International Conference on Pattern Recognition, ICPR 2016, Dec 2016, Cancún, Mexico. pp.2252 - 2257, 10.1109/ICPR.2016.7899971 . hal-01521628

HAL Id: hal-01521628
https://hal.archives-ouvertes.fr/hal-01521628
Submitted on 13 May 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Magnifying Subtle Facial Motions for 4D Expression Recognition

Qingkai Zhen  
IRIP Lab, School of Computer Science and Engineering, Beihang University, China, Email: qingkai.zhen@buaa.edu.cn

Di Huang  
IRIP Lab, School of Computer Science and Engineering, Beihang University, China, Email: dhuang@buaa.edu.cn

Yunhong Wang  
IRIP Lab, School of Computer Science and Engineering, Beihang University, China, Email: yhwang@buaa.edu.cn

Hassen Drira  
Institut Mines-Télécom/Télécom Lille, CRIStAL (UMR CNRS 9189), France  
Email: hassen.drira@telecom-lille.fr

Boulbaba Ben Amor  
Institut Mines-Télécom/Télécom Lille, CRIStAL (UMR CNRS 9189), France  
Email: boulbaba.benamor@telecom-lille.fr

Mohamed Daoudi  
Institut Mines-Télécom/Télécom Lille, CRIStAL (UMR CNRS 9189), France  
Email: mohamed.daoudi@telecom-lille.fr

Abstract—In this paper, an effective method is proposed to address the problem of automatic 4D Facial Expression Recognition (4D FER). The flow of 3D faces is first modeled to capture the spatial deformations based on the recently-developed Riemannian approach, namely Dense Scalar Fields (DSF), where registration and comparison of neighboring 3D face frames are jointly led. The deformations obtained are then fed into a temporal filtering based magnification step in order to amplify the slight facial actions over time. The proposed method allows revealing subtle (hidden) deformations which enhance the performance of emotion classification. We evaluate our approach on the BU4DFE dataset, and the state-of-art accuracy up to 94.18% is achieved, which is superior to the top one so far reported, clearly demonstrating its effectiveness.

I. INTRODUCTION

Facial expressions analysis and recognition from 3D data has attracted lots of researchers due to its diverse applications in the past decade, such as facial animation, human-computer interaction, etc. In recent years, there has been tremendous interest in tracking and recognition facial expressions from dynamic 3D facial expression sequence (4D data), it is suggested that the dynamics of facial expressions provides important cues about the underlying emotions that are not available in static 3D images.

There are a few works that use 4D data for facial expression analysis. Sun and Yin, the pioneers of 4D FER, extracted a Spatio-Temporal (ST) descriptor from dynamic sequences of 3D facial scans [1], and applied HMM classifier to predict the expression type. In [2], the tracking-model-based approach is presented for vertex correspondences, vertex motion estimation, and HMM is trained to learn the spatial and temporal information of the 3D model sequence. Canavan et al. [3] presented a dynamic curvature based approach for facial activity analysis, then constructed the dynamic curvature descriptor from local regions as well as temporal domains, and SVM classifier is adopted for classification. Sandbach et al. [4] exploited 3D motion-based features (Free-Form Deformation, FFD) between neighboring 3D facial geometry frames for FER. A feature selection step was applied to localize the features of each of the onset and offset segments of the expression. The HMM classifier was used to model the full temporal dynamics of each expression. In [5], the entire expressive sequence is modelled to contain an Onset followed by an Apex and an Offset. Feature selection methods are applied to extract features for each of the onset and offset segments of the expression. These features are then used to train GentleBoost classifiers and build an HMM to model the full temporal dynamics of the expression. Ben Amor et al [6], [7], [8] presented the facial expression deformation by collections of radial curves, Dense Scalar Fields (DSFs) features are feed into Random Forest or HMM classifier for classification. Xue et al. [9] applied three dimension discrete cosine transform (3D-DCT) on local depth patch-sequences to extract spatio-temporal features, and selected nearest-neighbor classifier to make decision.

Even though the performance of 4D FER has been great boosted in recent years, amplify the subtle movement on the facial surface is still an unsolved problem. We present a novel and effective approach to handle this problem, our contributions are two-folds:

- A comprehensive pipeline of spatio-temporal processing for effective facial expression recognition from 4D data.
- A method to amplify subtle movements on facial surfaces which contributes to distinguish similar expressions.

The rest of the paper is structured as follows. Section II introduces the background of the DSF based geometry feature. The magnification of subtle facial deformation is described in Section III. The experimental results are presented and analyzed in Section IV, followed by Section V where we conclude the paper.

II. DENSE SCALAR FIELDS

Following the geometric approach recently-developed in [6], we represent 3D facial surfaces by a collection of radial curves emanating from the nose tip. It is a parameterization imposed for 3D face description, registration, comparison, etc. The amount of deformation from one shape to another
Specifically, in the pre-processing step, the 3D face mesh in each frame is first aligned to the first one. The facial surfaces are then approximated by indexed collections of radial curves $\beta_\alpha$, where the index $\alpha$ denotes the angle formed by the curve with respect to a reference one. These curves are further uniformly resampled. Given a radial curve $\beta$ of the facial surface with an arbitrary orientation $\alpha \in [0, 2\pi]$, it can be parameterized as $\beta : I \rightarrow \mathbb{R}^3$, with $I = [0, 1]$, and mathematically represented using the square-root velocity function (SRVF), denoted by $q(t)$, according to:

$$q(t) = \frac{\dot{\beta}(t)}{\sqrt{\|\dot{\beta}(t)\|}}, t \in I. \quad (1)$$

This geometry representation has the advantage of capturing the curve shape and makes the calculus simpler. While there are several ways to measure the curve shape, an elastic analysis of the parametrized curves is appropriate in this application, particularly under facial expression variations. This is because (1) such analysis uses the square-root velocity function representation which allows for the comparison of local facial shapes in the presence of deformations; (2) this method employs a square-root representation under which the elastic metric is reduced to the standard $L^2$ metric and thus simplifies the analysis; (3) based on this metric, the group of re-parametrization acts by isometry on the curves manifold, and a Riemannian re-parametrization metric can thus be set between two facial curves. Shown in Fig. 1 are examples of apex frames taken from the 3D videos on the BU-4DFE dataset as well as the dense 3D deformations computed with respect to their neutral frames. Let us define the space of the SRVF's as

$$C = \{q : I \rightarrow \mathbb{R}^3, \|q\| = 1\} \subset L^2(I, \mathbb{R}^3) \quad (2)$$

where $\| \cdot \|$ indicates the $L^2$ norm. With the $L^2$ metric on its tangent space, $C$ becomes a Riemannian manifold. Basically, using this parametrization, each radial curve is represented on the manifold $C$ by its SRVF. Accordingly, given SRVF $q_1$ and $q_2$ of two curves, the shortest path $\psi^*$ on the manifold $C$ between them (called geodesic path) is a critical point of the following energy function:

$$E(\psi) = \frac{1}{2} \int \|\dot{\psi}(\tau)\|^2 d\tau \quad (3)$$

where $\psi$ denotes a path on the manifold $C$ between $q_1$ and $q_2$, $\tau$ is the parameter for traveling along the path $\psi$, and $\dot{\psi}(\tau) \in T_{\psi(\tau)}C$ is the tangent vector field on the curve $\psi(\tau) \in C$. Since elements of $C$ have a unit $L^2$ norm, $C$ is an hypersphere in the Hilbert space $L^2(I, \mathbb{R}^3)$. As a consequence, the geodesic path between any two points $q_1$ and $q_2 \in C$ is given by the minor arc of the great circle connecting them. The tangent vector field on this geodesic between the curves $\beta_1$ and $\beta_2$ making the angle $\alpha$ with the reference curve is parallel along the geodesic and one can represent it with the initial velocity vector (called shooting vector) without any loss of information.

$$\frac{d\psi^*_\alpha}{d\tau} \big|_{\tau=0} = \frac{\theta}{\sin(\theta)} (q_2 - \cos(\theta)q_1), \ (\theta \neq 0). \quad (4)$$

where $\theta = d_C(q_1, q_2) = \cos^{-1}(\langle q_1, q_2 \rangle)$ represents the length of the geodesic path connecting $q_1$ to $q_2$. In practice, the curves are re-sampled to a specified number of points, say $T$, and the face is approximated by a collection of $|A|$ curves. The norm of the quantity at each discrete point $r$ is computed to measure the amount of 3D deformation in this position of the surface parameterized by the pair $(\alpha, r)$, termed Dense Scalar Fields (DSFs). The final feature vector is of the size $T \times |A|$. We will refer to this quantity at a given time $t$ of the 3D video by $\chi(t)$ (see bottom row of Fig. 1 for illustration). It provides the amplitude of the deformations between two facial surfaces in a dense way.

### III. Subtle Facial Deformation Magnification

As described in Section II, $\chi$ reveals the shape difference between two facial surfaces by deforming one mesh to another through an accurate registration step. However, there exists another challenge to capture certain facial movements, especially the slight ones, with low spatial amplitude, reflected by the limited performance in distinguishing similar 3D facial expressions in the literature. To solve this problem, we propose a novel approach to highlight the subtle geometry changes of the facial surface in $\chi$ by adapting the Eulerian spatio-temporal processing [10] to the 3D domain.

The Eulerian spatio-temporal processing was introduced for motion magnification in 2D videos [10]. Its basic idea is to amplify the variation of pixel values over time in a spatially-multiscale manner without explicitly estimating motion but exaggerating it by amplifying temporal color changes at fixed
For the translational motion of the facial mesh, we express the observed \( \chi(s, t) \) value with respect to a displacement function \( \delta(t) \), such that \( \chi(s, t) = \chi(s) + \delta(t) \) and \( \chi(s, 0) = \chi(s) \). By using a first-order Taylor series expansion, at time \( t \), \( \chi(s + \delta(t)) \) can be approximated as

\[
\chi(s, t) \approx \chi(s) + \delta(t) \frac{\partial \chi(s)}{\partial s}
\]  
(5)

Let \( \phi(s, t) \) be the result of applying a broadband temporal band pass filter to \( \chi(s) \) at each position \( s \). Assume that the motion signal \( \delta(t) \) is within the pass band of the temporal filter.

\[
\phi(s, t) = \delta(t) \frac{\partial \chi(s)}{\partial s}
\]  
(6)

Amplify the band pass signal by factor \( \zeta \) and add it back to \( \chi(s) \).

\[
\hat{\chi}(s, t) = \chi(s, t) + \zeta \phi(s, t)
\]  
(7)

By combining (5), (6), and (7), we reach

\[
\hat{\chi}(s, t) \approx \chi(s) + (1 + \zeta(t)) \frac{\partial \chi(s)}{\partial s}
\]  
(8)

The process is shown in Fig. 2(a). Specifically, the video sequences are first decomposed into different spatial frequency bands by Gaussian pyramid, and these bands tend to be magnified differently. We consider that the time series correspond to the values of \( \chi \) on the mesh surfaces in a frequency band and apply a band pass filter to extract the frequency bands of interest. The temporal processing, \( \Sigma \), is uniform for all spatial levels and all \( \chi \) within each level. We then multiply the extracted band pass signal by a magnification factor \( \zeta \), and add the magnified signal to the original and collapse the spatial pyramid to obtain the final output.

The translational motion of the facial mesh, we express the observed \( \chi(s, t) \) value with respect to a displacement function \( \delta(t) \), such that \( \chi(s, t) = \chi(s) + \delta(t) \) and \( \chi(s, 0) = \chi(s) \). By using a first-order Taylor series expansion, at time \( t \), \( \chi(s + \delta(t)) \) can be approximated as

\[
\chi(s, t) \approx \chi(s) + \delta(t) \frac{\partial \chi(s)}{\partial s}
\]  
(5)

Let \( \phi(s, t) \) be the result of applying a broadband temporal band pass filter to \( \chi(s) \) at each position \( s \). Assume that the motion signal \( \delta(t) \) is within the pass band of the temporal filter.

\[
\phi(s, t) = \delta(t) \frac{\partial \chi(s)}{\partial s}
\]  
(6)

Amplify the band pass signal by factor \( \zeta \) and add it back to \( \chi(s) \).

\[
\hat{\chi}(s, t) = \chi(s, t) + \zeta \phi(s, t)
\]  
(7)

By combining (5), (6), and (7), we reach

\[
\hat{\chi}(s, t) \approx \chi(s) + (1 + \zeta(t)) \frac{\partial \chi(s)}{\partial s}
\]  
(8)

Assuming that the first-order Taylor expansion holds for the amplified larger perturbation \( (1 + \zeta) \delta(t) \), the motion magnification of 3D face video can be simplified as:

\[
\hat{\chi}(s, t) \approx \chi(s + (1 + \zeta) \delta(t))
\]  
(9)

This shows that the spatial displacement \( \delta(t) \) of the \( \chi(s) \) at time \( t \), is amplified to a magnitude of \( (1 + \zeta) \). Sometimes \( \delta(t) \) is not entirely within the pass band of the temporal filter. In this case, let \( \delta_k(t) \), indexed by \( k \), represent different temporal spectral components of \( \delta(t) \). The result in a band pass signal is,

\[
\phi(s, t) = \sum_k \gamma_k \delta_k(t) \frac{\partial \chi(s)}{\partial s}
\]  
(10)

where \( \gamma \) is an attenuation factor. Temporal frequency dependent attenuation can be equivalently interpreted as a frequency-
dependent motion magnification factor, $\zeta_k = \gamma \zeta$, and the amplified output signal is computed by

$$\hat{\chi}(s, t) \approx \chi(s + \sum_k (1 + \zeta_k) \delta_k(t))$$ (11)

Fig. 2(b) displays an example of facial deformation trajectory before (green) and after (red) magnification.

IV. EXPERIMENTAL RESULTS
A. Dataset and Protocol
The BU-4DFE dataset [1] is a dynamic 3D facial expression dataset which consists of 3D facial sequences of 58 females and 43 males. It includes in total 606 3D sequences possessing the 6 universal expressions. Each 3D sequence captures a facial expression at a rate of 25 fps (frames per second) and lasts approximately 3-4 seconds.

In our experiments, at a time $t$, the 3D face model $f^t$ is approximated by a set of 200 elastic radial curves originating from the nose tip, and a total of 50 points on each curve are sampled. Based on this parameterization, each 3D face geometry in the video sequence is compared to a reference frame $f^0$ to derive $\chi(t)$ at time $t$. Then, within the spatial processing step, a Gaussian pyramid decomposition is used to decompose $\chi$ into 4 band levels. Finally, a temporal processing to all the bands is applied. The factor $\zeta$ is set to 10, the sample rate $\zeta$ is set to 25, $f \in [0.3, 0.4]$, and the attenuation rate $\gamma$ is set to 1.

It should be noted that the proposed approach can either be evaluated when making use of full sequences [6], [9], [11] or sliding window-based sub-sequences [1], [2], [5], [12], while as pointed in [13], the latter can bias the final result. As a result, we exploit the former, and our experiments are conducted on two sub-pipelines: (1) the whole video sequence (denoted as WV) and (2) the magnified whole video sequence (denoted as MWV).

A multi-class Support Vector Machine (SVM) is exploited where $\hat{\chi}$ is treated as a feature vector to predict the expression label. We also adopt HMM to encode the temporal behavior of the sequence for decision. To allow fair comparison with the previous studies, we randomly select 60 subjects from the BU-4DFE dataset under 10-fold cross-validation protocol.

B. Performance
Table I shows that the magnification procedure achieves an improvement of around 10% in the accuracy for both classifiers, i.e. SVM and HMM. Without magnification, our approach reaches the performance of 82.49% and 83.19% using the SVM and HMM classifiers, while the results are improved to 93.39% and 94.18%, which highlights the effectiveness of magnification in 3D face videos. Table II shows the confusion matrices (WV, MWV) achieved by using the SVM and HMM classifiers. From this table, we can see that the SU and HA sequences are better predicted than the others. This is mainly due to the clear patterns and high intensities of their deformations. The remaining expressions (DI, FE, AN and SA) are harder to distinguish. We believe that two major reasons induce this difficulty: (1) intra-class variations make similar classes confusing, such as DI/AN/FE; and (2) lower deformation magnitude is often exhibited when these expressions are performed. Furthermore, it can be seen from these confusion matrices, the accuracies in distinguishing AN, DI and FE expressions are all significantly ameliorated. Fig. 3 gives more illustrations of deformation magnification on the sequences of the same subject possessing the six prototype expressions.

C. Comparison with state-of-the-art
Several studies report their FER results on the BU-4DFE dataset; however, they differ in the experimental setting. In this section, we compare our results with the one of the existing approaches considering these differences.

Top results on BU-4DFE are shown in Table III. In this table, #E denotes the number of expressions, #S is the number of subjects, #CV provides the number of cross-validation, and Full Seq./Win means the decision is made based on full sequence or sub-sequences captured using a sliding window. [2] reports the highest accuracy when using a sliding window of 6 frames; nevertheless, the approach requires manual annotation of 83 landmarks on the first frame. Moreover, the vertex-level dense tracking is time consuming. In a more recent work from the same group developed by Reale et al. [12], the authors deliver a classification rate of 76.9% using the sequences of 100 subjects with a fixed size of window of 15 frames, when segmentation is manually applied to the 3D face video to extract the expressive time interval. In [14], Fang et al. reach an accuracy of 74.63% with 507 sequences of 100 subjects, but they do not provide details on the classification scheme. Le et al. [11] evaluate their algorithm on the sequences of 60 subjects only on three expressions (HA, SA and SU) and display the result of 92.22%. Regarding on the tasks that conduct classification under the same protocol [6], [9], [13], the proposed method outperforms them, demonstrating its competency at 4D FER. Besides, it also possesses the advantages: (1) no manual landmark is required; and (2) no dimensionality reduction or feature selection techniques are applied.

V. CONCLUSIONS
In this paper, an effective approach is presented for 4D FER. It focuses on improving the performance by 3D video magnification which reveals subtle facial deformations. After a preprocessing step, the flow of 3D faces is first modeled to capture spatial shape changes in the DSF based Riemannian geometry space, where registration and comparison are jointly achieved. Such deformations are then amplified using the temporal filter over time. The prediction is finally carried out using these magnified features. Experimental results on the BU-4DFE dataset demonstrate the effectiveness of the proposed method.
TABLE I: Average accuracy with standard deviation achieved by SVM and HMM using full sequence before and after magnification on the BU-4DFE database.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Magnification?</th>
<th>Performance (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM on ( \chi )</td>
<td>N</td>
<td>82.49 ± 3.10</td>
</tr>
<tr>
<td>SVM on ( \chi )</td>
<td>Y</td>
<td>93.39 ± 3.54</td>
</tr>
<tr>
<td>HMM on ( \chi(t) )</td>
<td>N</td>
<td>83.19 ± 2.83</td>
</tr>
<tr>
<td>HMM on ( \chi(t) )</td>
<td>Y</td>
<td>94.18 ± 2.46</td>
</tr>
</tbody>
</table>

TABLE II: Confusion matrices (WV, MWV) achieved by the SVM and HMM classifiers respectively on the BU-4DFE database.

<table>
<thead>
<tr>
<th>SVM on ( \chi )</th>
<th>Whole Video (WV)</th>
<th>Magnified Whole Video (MWV)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( % )</td>
<td>AN</td>
<td>DI</td>
</tr>
<tr>
<td>AN</td>
<td>73.86</td>
<td>9.18</td>
</tr>
<tr>
<td>DI</td>
<td>8.76</td>
<td>71.27</td>
</tr>
<tr>
<td>FE</td>
<td>5.79</td>
<td>5.37</td>
</tr>
<tr>
<td>HA</td>
<td>0.81</td>
<td>1.18</td>
</tr>
<tr>
<td>SA</td>
<td>2.54</td>
<td>2.27</td>
</tr>
<tr>
<td>SU</td>
<td>0.74</td>
<td>0.88</td>
</tr>
<tr>
<td>Average</td>
<td>82.49 ± 3.10</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>HMM on ( \chi(t) )</th>
<th>Whole Video (WV)</th>
<th>Magnified Whole Video (MWV)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( % )</td>
<td>AN</td>
<td>DI</td>
</tr>
<tr>
<td>AN</td>
<td>75.29</td>
<td>5.88</td>
</tr>
<tr>
<td>DI</td>
<td>10.42</td>
<td>71.55</td>
</tr>
<tr>
<td>FE</td>
<td>5.07</td>
<td>6.86</td>
</tr>
<tr>
<td>HA</td>
<td>0.48</td>
<td>0.87</td>
</tr>
<tr>
<td>SA</td>
<td>3.71</td>
<td>1.01</td>
</tr>
<tr>
<td>SU</td>
<td>0.49</td>
<td>0.33</td>
</tr>
<tr>
<td>Average</td>
<td>83.19 ± 2.83</td>
<td></td>
</tr>
</tbody>
</table>

TABLE III: Comparative results with the state-of-the-art on BU-4DFE.

<table>
<thead>
<tr>
<th>Method</th>
<th>Experimental Settings</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sun et al. [1]</td>
<td>6E, 60S, 10-CV, Win=6</td>
<td>90.44%</td>
</tr>
<tr>
<td>Sun et al. [2]</td>
<td>6E, 60S, 10-CV, Win=6</td>
<td>94.37%</td>
</tr>
<tr>
<td>Reale et al. [12]</td>
<td>6E, 100S, --, Win=15</td>
<td>76.9%</td>
</tr>
<tr>
<td>Sandb. et al. [5]</td>
<td>6E, 60S, 6-CV, Win</td>
<td>64.6%</td>
</tr>
<tr>
<td>Fang et al. [14]</td>
<td>6E, 100S, 10-CV, --</td>
<td>74.63%</td>
</tr>
<tr>
<td>Le et al. [11]</td>
<td>3E, 60S, 10-CV, Full seq.</td>
<td>92.22%</td>
</tr>
<tr>
<td>Xue et al. [9]</td>
<td>6E, 60S, 10-CV, Full seq.</td>
<td>78.8%</td>
</tr>
<tr>
<td>Berretti et al. [13]</td>
<td>6E, 60S, 10-CV, Full seq.</td>
<td>79.4%</td>
</tr>
<tr>
<td>Berretti et al. [13]</td>
<td>6E, 60S, 10-CV, Win=6</td>
<td>72.25%</td>
</tr>
<tr>
<td>Ben Amor et al. [6]</td>
<td>6E, 60S, 10-CV, Full seq.</td>
<td>93.21%</td>
</tr>
<tr>
<td>Ben Amor et al. [6]</td>
<td>6E, 60S, 10-CV, Win=6</td>
<td>93.83%</td>
</tr>
<tr>
<td>This work – SVM on ( \chi )</td>
<td>6E, 60S, 10-CV, Full seq.</td>
<td>93.59%</td>
</tr>
<tr>
<td>This work – HMM on ( \chi(t) )</td>
<td>6E, 60S, 10-CV, Full seq.</td>
<td>94.18%</td>
</tr>
</tbody>
</table>

REFERENCES
Fig. 3: Illustrations of the deformation magnification on the sequences of the same subject performing the six universal expressions. One can appreciate the magnification effects on 3D deformations compared to those of the original \textit{DSF} feature. From up to bottom, each row presents the texture image, the depth map, the original \textit{DSF} feature, and the amplified feature, respectively.