Magnifying Subtle Facial Motions for Effective 4D Expression Recognition

Qingkai Zhen, Di Huang, Hassen Drira, Boulbaba Ben Amor, Yunhong Wang, Mohamed Daoudi

Magnifying Subtle Facial Motions for Effective 4D Expression Recognition

Qingkai Zhen, Di Huang, *Member, IEEE*, Hassen Drira, Boulbaba Ben Amor, *Senior Member, IEEE*, Yunhong Wang, *Senior Member, IEEE*, and Mohamed Daoudi, *Senior Member, IEEE*

Abstract—In this paper, an effective approach is proposed for automatic 4D Facial Expression Recognition (*FER*). It combines two growing but disparate ideas in the domain of computer vision, *i.e.* computing spatial facial deformations using a Riemannian method and magnifying them by a temporal filtering technique. Key frames highly related to facial expressions are first extracted from a long 4D video through a spectral clustering process, forming the Onset-Apex-Offset flow. It is then analyzed to capture the spatial deformations based on Dense Scalar Fields (DSF), where registration and comparison of neighboring 3D faces are jointly led. The generated temporal evolution of these deformations is further fed into a magnification method to amplify facial activities over time. The proposed approach allows revealing subtle deformations and thus improves the emotion classification performance. Experiments are conducted on the BU-4DFE and BP-4D databases, and competitive results are achieved compared to the state-of-the-art.

Index Terms—4D Facial Expression Recognition, Riemannian geometry, subtle motion magnification, key frame detection.

1 Introduction

Facial expressions are important non-verbal ways for humans to communicate their feelings and emotion states. In recent years, as a major topic of affective computing, Facial Expression Recognition (*FER*) has received increasing interests from various communities due to its wide potential in many applications [1], [2], such as psychological analysis, transport security (*e.g.* driver fatigue alerting), human-machine interaction, animation of 3D avatars, *etc*. In particular, *FER* is expected to make crucial contributions to the next-generation interaction system [3]. On the other side, *FER* is a very challenging issue mainly because of intra-class diversity and complexity of imaging environments [4]. In 1970s, Ekman and Friesen [5] define six types of universal facial expressions, *i.e.* Anger (*AN*), Disgust (*DI*), Fear (*FE*), Happiness (*HA*), Sadness (*SA*), and Surprise (*SU*), which are consistent across diverse sexes, races and cultures. Inspired by such a fundamental, *FER* research is extensively developed and a large number of *FER* approaches are proposed. They can be roughly categorized into four classes according to the imaging channel – (1) *FER* from 2D still images or video streams [6]–[8]; (2) *FER* from 3D static scans [9]–[11]; (3) *FER* from 4D (or 3D+*Time*) dynamic flow of facial scans [12]–[14]; and 4) *FER* from 2D and 3D data combination [4], [15].

Despite the great progress made in 2D-based *FER* approaches, they suffer from the problem of illumination and pose variations, which often occur in real-life conditions. With the rapid innovation of devices in 3D data acquisition, the 3D technique is regarded as a promising alternative to achieve robust *FER*, since 3D data directly reflect deformations of facial surfaces caused by muscles movement [16], highly related to facial expressions. Additionally, they present immunity to the changes in lighting and viewpoint. More recently, the advent of 4D imaging systems makes it possible to deliver 3D shape sequences of high quality for more comprehensive facial expression analysis. Besides the geometric attributes in each frame (static 3D scan), the 3D scan sequence also captures the quasi periodical dynamic variations of facial expressions from adjacent frames. In 3D and 4D *FER*, the most important step is to represent shape and deformation patterns of different expressions, where the features produced are expected to possess good distinctiveness to describe the expressions in a person-independent manner. In 3D *FER*, the global or local information extracted from static 3D facial scans, encoded in certain features, is directly fed into classifiers for decision making. In 4D *FER*, additional features are also required to convey dynamic changes of adjacent 3D facial scan frames, which are adopted together with the ones in 3D for classification.

Although *FER* performance has been substantially boosted by 4D data in recent years, there still exist some unsolved problems. On the one hand, some reputed similar expressions are difficult to distinguish since their facial deformations are sometimes really slight [11]. On the other hand, detection of key frames that include expressions information in the 3D face video is not paid sufficient attention. To deal with these problems, this paper presents a novel and effective approach to automatic 4D *FER*. First, it addresses the issue of capturing subtle deformations in 4D videos. Inspired by the achievements of motion magnification in 2D videos [17]–[20], we propose a method to highlight subtle facial motions in the 4D domain. The main difficulty is to establish a vertex-level dense correspondence between frames. This point is solved by previous Dense Scalar Fields (*DSFs*) computation [14], where an accurate registration of neighboring faces is achieved through an elastic matching of radial facial curves. Then, based on Eulerian spatio-temporal processing, facial motions are magnified especially the subtle ones and deformations of certain expressions with low intensities are amplified, leading to improved classification accuracy. Meanwhile, a *Deformation Spectral Clustering* (*DSC*) based key-frame detection algorithm is adopted to locate the Onset-Apex-Offset states of each expression. The idea behind is to segment the frames when the expression happens in the video by analyzing the

Qingkai Zhen, Di Huang and Yunhong Wang are with the Laboratory of Intelligent Recognition and Image Processing, School of Computer Science and Engineering, Beihang University, Beijing 100191, China (Email: {qingkai.zhen, dhuang, yhwang}@buaa.edu.cn).
B. Ben Amor, H. Drira, and M. Daoudi are with IMT Lille Douai, Univ. Lille, CNRS, UMR 9189 - CRIStAL, Lille F-59000, France.

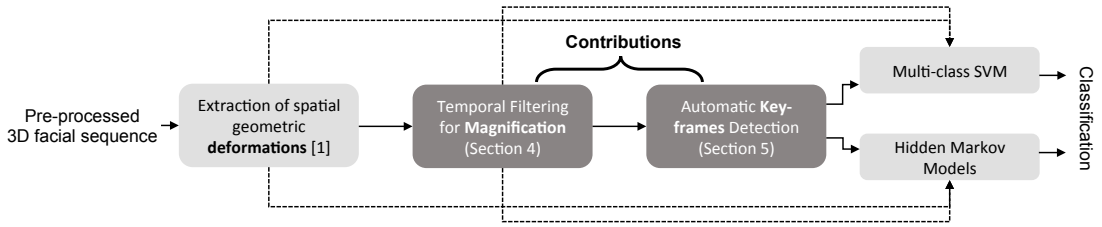


Fig. 1: Overview of the proposed pipeline for 4D FER. From left-to-right: **Extraction of geometric (deformation) features termed *DSFs* – Magnification of facial motions – Key-frames (*i.e.* Onset-Apex-Offset) detection – Classification technique: SVM or HMM**.

consistency of the shape deformation dynamics. Two classification techniques, *SVM* and *HMMs*, are finally employed for label prediction. To validate the proposed approach, extensive experiments are conducted on the BU-4DFE and BP-4D databases. The results are very competitive, clearly demonstrating its competency, brought by (i) the magnification step, and (ii) the key-frame detection step. In summary, the main contributions of this study are,

- A novel and effective approach is presented for 4D *FER*, which detects key frames and highlight geometric dynamics in predicting expression labels. The pipeline is depicted in Fig.1.
- A magnification method is introduced to highlight subtle movements on facial surfaces which contributes to distinguishing similar expressions.
- A consistency-based method is proposed to localize key-frames of 3D facial scan sequences, which is able to find the key states of individual expressions in videos.
- State of the art results are achieved on the BU-4DFE and BP-4D datasets compared to the ones in literature, without use of landmarks and dimensionality reduction or feature selection techniques.

Some preliminary results of the motion magnification method are demonstrated in our previous work [21]. This study includes those results but significantly extends them in the following ways. Firstly, we discuss the related work more comprehensively, which helps to highlight the advantage of the proposed approach. Secondly, we propose the *DSC* method to segment the frames when expressions appear, and it makes the whole pipeline more efficient and practical. Thirdly, we not only present additional experiments on BP-4D to confirm the effectiveness of the proposed approach, but also show more analysis on them. Lastly, we reach better final performance on the BU-4DFE database.

The remainder of the paper is structured as follows. In Section 2, a brief review of existing approaches in 3D FER and 4D FER is given. The background on the Riemannian method for 3D face analysis and the derived Dense Scalar Fields (*DSFs*) is presented in Section 3. The magnification of subtle facial deformations is introduced in Section 4, and Section 5 describes the Deformation Spectral Clustering (*DSC*) algorithm for key-frames detection. The comprehensive experimental study is displayed and discussed in Section 6, followed by Section 7 with the concluding remarks and future directions.

2 Related Work

As we state above, 4D face data contain static cues of individual 3D facial scans and dynamic cues between adjacent ones, and 4D *FER* thus has to deal with two main issues, *i.e.* shape representation for static 3D facial scans and dynamic description in facial scan sequences. In this section, we review the techniques of curve based 3D facial representation as well as the methods in 4D *FER*.

2.1 Curve based 3D Facial Representation

The task of representing 3D static facial surfaces (in the form of pointclouds, meshes, range images, *etc.*) has been approached in many ways, leading to a serial of successes. In general, the most popular techniques go into the categories of template based, local region or feature based, and curve based.

Template-based methods fit a rigid or deformable generic face model to input scans under certain criteria, and the coefficients achieved are regarded as expression related features in classification. Bilinear Models [22], *Annotated Face Model* (*AFM*) [23], and *Statistic Facial feAture Model* (*SFAM*) [24] are some typical examples. Local region or feature based methods extract various geometric and topological properties from a number of selected areas of 3D faces, and quantizations of such properties are adopted to describe expressions for prediction. Representatives include the low level geometric feature pool [25], *Local Normal Patterns* (*LNP*) [26], histograms of differential geometry quantities [4], *etc*.

Compared to the two types of methods aforementioned, curve-based methods appear more recently. They compute deformations of expressions by measuring shapes of a set of sampling curves in the Riemannian space. In [27], Samir *et al.* propose the framework to represent facial surfaces by an indexed collection of 3D closed curves. These curves are level curves of a surface distance function defined as the length of the shortest path between the fixed nose tip point and a point of the extracted curve along the facial surface. It enables comparing 3D facial shapes by corresponding curves. In [9], Maalej *et al.* implement the curve based idea to 3D *FER* and compute the length of the geodesic path between corresponding patches of 3D faces. It provides quantitative information between facial surfaces of different expressions. The curve based method proves effective in 3D shape description, and is being applied to more and more fields, such as 3D face recognition [28], 3D facial gender classification [29], 3D facial age estimation [30], 3D action recognition [31], *etc*.

2.2 Advances in 4D *FER*

Regarding 4D *FER*, this topic has been widely investigated in the past decade with its performance consistently increasing. Besides static facial shape representation, dynamic feature modelling plays a crucial role in 4D *FER* as well.

Sun and Yin, the pioneers of 4D *FER*, calculate the change of a generic deformable face model to extract a *Spatio-Temporal* (*ST*)

descriptor from the sequence of 3D scans [32]. The vertex flow tracking is applied to each frame to form a set of motion trajectories of the 3D face video. The *ST* features and *HMM* are used for classification. A similar idea appears in [33], where Canavan *et al.* describe dynamics of 3D facial surfaces by curvature-based shape-index values. A 3D face tracking model detects the local regions across 3D dynamic sequences, and the features are characterized in the regions along the temporal axis. *Linear Discriminant Analysis* (*LDA*) is employed for dimension reduction and *Support Vector Machine* (*SVM*) is adopted to predict the expression label.

Sandbach *et al.* [34] exploit 3D motion-based features *Free-Form Deformation* (*FFD*) between neighboring 3D facial geometry frames for *FER*. A feature selection step is applied to localize the features of each onset and offset segment of the expression. *HMM* is used to model the temporal dynamics of each expression. In another work of the same authors [3], an expression is modeled to a sequence which contains an onset, an apex and an offset. The features selected are utilized to train *GentleBoost* and build *HMM* to classify the expressive sample.

Fang *et al.* [10], [13] highlight facial shape registration and correspondence between 3D meshes in videos along the temporal line, and a variant of *Local Binary Patters* (*LBP*), namely *LBP on Three Orthogonal Plane* (*LBP-TOP*), is introduced for dense local feature based static and dynamic facial representation. The *SVM* classifier with a *Radial Basis Function* (*RBF*) kernel is employed to distinguish expressions.

Berretti *et al.* [35] present an automatic and real-time approach to 4D *FER*. It also makes use of local shape features but in a sparse manner. A set of 3D facial landmarks are firstly detected, and local characteristics of the facial patches around those landmarks and their mutual distances are utilized to model facial deformations. *HMM* is adopted in expression classification. Similarly, Xue *et al.* [36] crop local depth patch-sequences from consecutive expression frames based on automatically detected facial landmarks. Three-dimensional *Discrete Cosine Transform* (*3D-DCT*) is then applied on these shape patch-sequence to extract spatio-temporal features for dynamic facial expression representation. After feature selection and dimension reduction, the features are finally fed into the nearest-neighbor classifier.

Reale *et al.* [37] propose a shape feature, called *Nebula* to improve the performance of *FER* and *Action Unit* (*AU*) recognition. For a given spatio-temporal volume, the data is voxelized and fit to a cubic polynomial. The labels and angles calculated according to the principal and least curvatures are used to construct a histogram for each face region, and the one by concatenating those of all the regions forms the final vector. Key frames are manually intervened in 3D face videos, related to the most intense expression, and their features are regarded as the input of (*SVM*).

Hayat and Bennamoun [38] calculate local video shape patches of variable lengths from numerous locations and represent them as the points on the Grassmannian manifold. A graph-based spectral clustering algorithm is exploited to separately cluster these points for every individual expression class. Using a valid Grassmannian kernel function, the resulting cluster centers are embedded into a *Reproducing Kernel Hilbert Space* (*RKHS*) where six binary *SVM* models are learned for classification.

In [14], [39], [40], Ben Amor *et al.* represent facial surfaces by different collections of radial curves. Riemannian shape analysis is carried out to quantify dense facial shape deformations and extract motion cues from 3D frame sequences. Two different classification schema are successively performed, *i.e.* an *HMM-based* classifier and a mean deformation-based classifier. These studies emphasize the capability of the proposed geometric deformation analysis to recognize expressions in 3D videos.

3 Background on Dense Scalar Fields

Following the geometric approach recently-developed in [14], we represent 3D facial surfaces by collections of radial curves which emanate from nose tips. It is a new parameterization imposed for 3D face description, registration, comparison *etc*. The amount of deformation from one shape into another (across the 3D video) is computed through analyzing shapes of 3D radial curves using the tools of differential geometry.

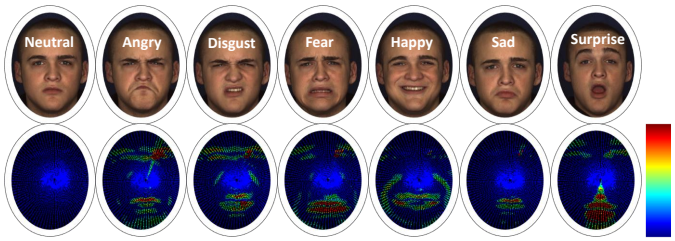


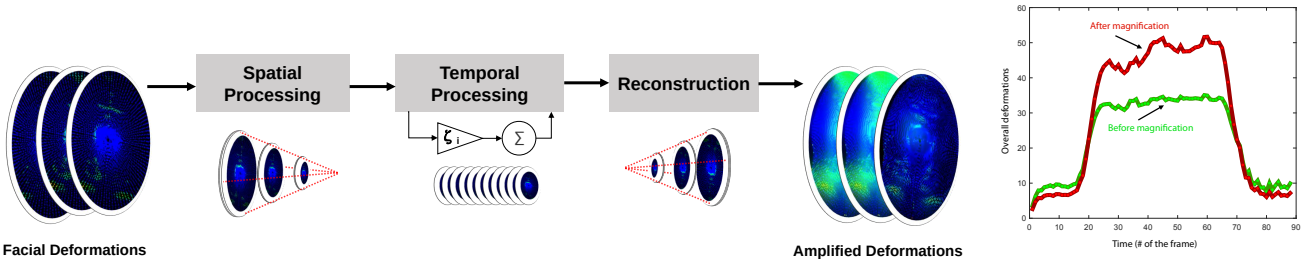
Fig. 2: Top row: facial texture images of an individual possessing different expressions. Bottom row: facial deformations captured by *DSFs*, where warm colors are associated to the facial regions with high deformations while cold colors reflect the most static areas.

In the pre-processing step, the 3D mesh in a frame is aligned to the first mesh and cropped for pose normalization and informative part segmentation. Each facial surface is then approximated by an indexed collection of radial curves, termed β_α, where the index, α, denotes the angle formed by the curve with respect to a reference (*i.e.* $\alpha = 0$). The curves β_α are further uniformally resampled. To simplify the notation, we denote by β a curve on the surface. Given a radial curve β of the face with an arbitrary orientation $\alpha \in [0, 2\pi]$, it can be parameterized as $\beta : I \to \mathbb{R}^3$, with $I = [0, 1]$, and mathematically represented by using the *Square-Root Velocity Function* (*SRVF*), denoted by $q(t)$, according to:

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

This special representation has the advantage of capturing the curve shape in a parametrization-invariant manner. While there are several ways to analyze shapes of curves, an elastic analysis of the parametrized curves is particularly appropriate to achieve accurate registration across neighboring 3D facial scans. This is because (1) such analysis adopts the *SRVF* representation which allows us to compare local facial shapes in the presence of deformations, (2) the elastic metric is substituted to the standard $\mathbb{L}^2$ metric and thus simplifies the analysis, and (3) under this metric the group of re-parametrization acts by isometry on the curve manifold, and a Riemannian re-parametrization metric can thus be set between two facial curves. Shown in Fig. 2 are examples of apex frames taken from the 3D videos of the BU-4DFE dataset as well as the dense 3D deformations computed with respect to the neutral frame. Let us define the space of the *SRVFs* as

$$\mathcal{C} = \{\mathbf{q} : I \to \mathbb{R}^3, \|\mathbf{q}\| = 1\} \subset \mathbb{L}^2(I, \mathbb{R}^3), \tag{2}$$



(a) (b)

Fig. 3: (a) Overview of 3D video magnification. The original facial deformation features are first decomposed into different spatial frequencies, and the temporal filter is applied to all the frequency bands. The filtered spatial bands are then amplified by a given factor ζ, added back to the original signal, and collapsed to the output sequence. (b) An example of facial expression deformation (norm of the velocity vector) before (green) and after (red) magnification.

where $\|\cdot\|$ indicates the $\mathbb{L}^2$ norm. With the $\mathbb{L}^2$ metric on its tangent space, $\mathcal{C}$ becomes a Riemannian manifold. Basically, based on this parametrization, each radial curve is represented on the manifold $\mathcal{C}$ by its *SRVF*. According to Eqn. (1), given *SRVFs* $\mathbf{q}_1$ and $\mathbf{q}_2$ of two curves β_1 and β_2, the shortest path ψ^* connecting them on the manifold $\mathcal{C}$ (also called geodesic path) is a critical point of the following energy function:

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

where ψ denotes the path on the manifold $\mathcal{C}$ between $\mathbf{q}_1$ and $\mathbf{q}_2$, and τ is the parameter for travelling along path ψ. The quantity $\dot{\psi}(\tau) \in T_{\psi(\tau)}(\mathcal{C})$ is the tangent vector field on the curve $\psi(\tau) \in \mathcal{C}$. Since elements of $\mathcal{C}$ have a unit $\mathbb{L}^2$ norm, $\mathcal{C}$ is an hypersphere in the Hilbert space $\mathbb{L}^2(I, \mathbb{R}^3)$. As a consequence, the geodesic path between any two points $\mathbf{q}_1$ and $\mathbf{q}_2 \in \mathcal{C}$ is given by the minor arc of the great circle connecting them. The tangent vector field on this geodesic between any $\mathbf{q}_1$ and $\mathbf{q}_2$ is parallel along the geodesic and one can represent it by the initial velocity vector (*i.e.* shooting vector), without any loss of information as follows,

$$\frac{d\psi^*}{d\tau}_{|_{\tau=0}} = \frac{\theta}{\sin(\theta)}(\mathbf{q}_2 - \cos(\theta)\mathbf{q}_1), \ (\theta \neq 0). \tag{4}$$

where $\theta = d_{\mathcal{C}}(\mathbf{q}_1, \mathbf{q}_2) = cos^{-1}(\langle \mathbf{q}_1, \mathbf{q}_2 \rangle)$ represents the length of the geodesic path connecting $\mathbf{q}_1$ to $\mathbf{q}_2$.

Fig.4 shows an interpretation of this representation. In Fig.4 (a), a sample face of a happy expression and its radial curves are given. In Fig.4 (b) and (c), two corresponding radial curves on a neutral and a happy face of the same person are highlighted. The two curves are plotted together in Fig.4 (d), and the amount of deformation between them is calculated using Eqn. (4) with the magnitude of this vector depicted in Fig.4 (e). Fig.4 (f) illustrates the way to map the two facial radial curves on the hypersphere $\mathcal{C}$ in the Hilbert space through their *SRVFs* $\mathbf{q}_1$ and $\mathbf{q}_2$ and displays the geodesic path connecting them.

In practice, the facial surface is approximated by a collection of $|\Lambda|$ curves, and the curves are re-sampled to a discrete number of points, $h \in [1, H]$. The norm of the quantity at h is calculated to represent the amount of 3D shape deformation at this point on the surface parameterized by the pair (α, h), termed *Dense Scalar Fields* (*DSFs*). The final feature vector is thus of size $H \times |\Lambda|$. We refer to this quantity at a given time t of the 3D face video by $\chi(t)$ (see the bottom row of Fig.2 for an illustration). It provides the amplitude of the deformation between two facial surfaces in a dense way. Alg.1 describes the procedure of *DSF* computation on a current frame.

Algorithm 1: *DSF* Computation

Input:
F^0 and F^t: reference and current facial surfaces;
H: number of the sample points on a curve;
$\Delta\alpha$: angle between successive radial curves;
$|\Lambda|$: number of curves per face;
Output:
$\chi(t)$: DSF feature of F^t
Procedure: ComputeDSF(F^0, F^t, H, $\Delta\alpha$, $|\Lambda|$)
  $n \leftarrow 0$
  while $n < |\Lambda|$ **do**
    $\alpha = n \cdot \Delta\alpha$
    for $i \leftarrow 0, t$ **do**
      extract curve β^i_α
      compute *SRVF* of β^i_α:
        $\mathbf{q}^i_\alpha(h) \doteq \frac{\dot{\beta}^i_\alpha(h)}{\sqrt{\|\dot{\beta}^i_\alpha(h)\|}} \in \mathcal{C}, h = 1, ..., H$
    end for
    compute distance between $\mathbf{q}^0_\alpha$ and $\mathbf{q}^t_\alpha$
    $\theta = d_{\mathcal{C}}(\mathbf{q}^0_\alpha, \mathbf{q}^t_\alpha) = cos^{-1}(\langle \mathbf{q}^0_\alpha, \mathbf{q}^t_\alpha \rangle)$
    compute deformation vector $\frac{d\psi}{\tau}_{|_{\tau=0}}$ using Eqn.(4):
    $f(\mathbf{q}^0_\alpha, \mathbf{q}^t_\alpha) = (\chi_\alpha(1), \chi_\alpha(2), ..., \chi_\alpha(H)) \in R^H_+$
    $\chi_\alpha(h) = \left| \frac{\theta}{sin(\theta)}(\mathbf{q}^t_\alpha(h) - cos(\theta)\mathbf{q}^0_\alpha(h)) \right|$
    $h = 1, ..., H$
  end while
  compute $\chi(t)$ as the magnitude of $\frac{d\psi}{\tau}_{|_{\tau=0}}(t)$:
    $\chi(t) = (f(\mathbf{q}^0_0, \mathbf{q}^t_0), ..., f(\mathbf{q}^0_{|\Lambda|}, \mathbf{q}^t_{|\Lambda|}))$
  return $\chi(t)$
end procedure

4 Subtle Facial Deformation Magnification

As described in Section 3, χ reveals the shape difference between two facial surfaces by deforming one mesh into another through an accurate registration step. However, there exists another challenge to capture certain facial movements, in particular the slight ones, with low spatial amplitudes, 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 variations of the facial surface in χ by adapting

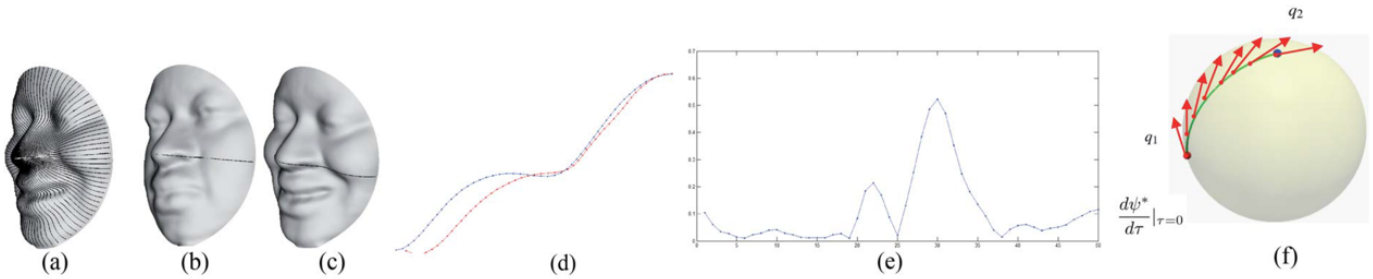


Fig. 4: *DSF* illustration: (a) Radial curves on a facial surface; (b) and (c) corresponding curves on a neutral face and a happy face of the same subject; (d) plotting curves together; (e) values of the magnitude computed between the curves in (d); and (f) parallel vector field across the geodesic between $\mathbf{q}_1$ and $\mathbf{q}_2$ on the hypersphere $\mathcal{C}$ in the Hilbert space.

Eulerian spatio-temporal processing [20] to the 3D domain. This method and its application to 3D face videos are presented in the subsequent.

Eulerian spatio-temporal processing was introduced for motion magnification in 2D videos and has shown its effectiveness in [20]. The basic idea is to highlight the change of pixel values over time, in a spatially-multiscale way, without explicitly estimating motion but exaggerating motion by amplifying temporal color changes at fixed positions. It depends on a linear approximation related to the assumption of brightness constancy which forms the fundamental theory of the optical flow algorithm. However, the case is not that straightforward in 3D, since the correspondence of vertices across frames cannot be achieved as easy as that in 2D. Fortunately, during computation of *DSFs*, such correspondence is established by surface registration along its radial curves. Therefore, we can adapt Eulerian spatio-temporal processing to the 3D case, where we take into account the value along the time axis at any spatial location and highlight the differences in a given temporal frequency band of interest, combining spatial and temporal processing to emphasize subtle changes in 3D face videos.

Fig.3(a) illustrates this process. Specifically, a video sequence is first decomposed into different spatial frequency bands by using Gaussian pyramid, and these bands might be magnified differently. The time series correspond to the values of χ on the mesh surface in the frequency band and a band pass filter is applied to extract the frequency bands of interest. Temporal processing, denoted as $\mathfrak{T}$, is uniform for all spatial levels and for all χ within each level. The band passed signal extracted is then multiplied by a magnification factor ζ, the magnified signal is added to the original, and the spatial pyramid is collapsed to obtain the final output.

Let $\chi(s,t)$ denote *DSF*, *i.e.* the amplitude of the deformations, between the facial surface at position s and time t and the predefined reference. Since *DSF* undergoes translational motion, we express $\chi(s,t)$ with respect to a displacement function $\delta(t)$, such that $\chi(s,t) = f(s+\delta(t))$ and $\chi(s,0) = f(s)$. The objective of motion magnification is to synthesize the signal as:

$$\hat{\chi}(s,t) = f(s+(1+\zeta)\delta(t)) \tag{5}$$

for a given amplification factor ζ.

By using a first-order Taylor series expansion, the *DSF* feature at time t, $f(s+\delta(t))$ is written as a first-order Taylor expansion about position s:

$$\chi(s,t) \approx f(s) + \delta(t)\frac{\partial f(s)}{\partial s} \tag{6}$$

Algorithm 2: Online 3D Deformation Magnification

Input: χ-*DSF*, l-Gaussian pyramid levels, ζ-amplification factor, ξ-sample rate, γ-attenuation rate, $\mathfrak{f}$-facial expression frequency, n number of frames

Step1. Spatial Processing

for $i = 1; i \leq n$ **do**

  $\mathfrak{D}(i,:,:,:)$ = decompose $\chi(i)$, with l level Gaussian pyramid.

Step2. Temporal Processing

$\mathfrak{S}$ = $\mathfrak{T}(\mathfrak{D}, \mathfrak{f}, \xi)$

Step3. Magnification (for different color channels)

for $i = 1; i \leq 3$ **do**

  $\mathfrak{S}(:,:,:,\mathfrak{i})$ = $\mathfrak{S}(:,:,:,\mathfrak{i}) * \zeta * \gamma$

Step4. Reconstruction

for $i = 1; i \leq n$ **do**

  $\hat{\chi}(i)$ = $\mathfrak{S}(i,:,:,:)$ + $\chi(i)$

Output: $\hat{\chi}(t)$

Let $\phi(s,t)$ be the result when applying a broadband temporal bandpass filter to $\chi(s,t)$ at each position s. We assume that the motion signal $\delta(t)$ is within the passband of the temporal bandpass filter, and we have

$$\phi(s,t) = \delta(t)\frac{\partial f(s)}{\partial s} \tag{7}$$

We amplify the bandpass signal by ζ and then add it back to $\chi(s,t)$, achieving

$$\tilde{\chi}(s,t) = \chi(s,t) + \zeta\phi(s,t) \tag{8}$$

By combining Eqn. (6), (7), and (8), we reach

$$\tilde{\chi}(s,t) \approx f(s) + (1+\zeta)\delta(t)\frac{\partial f(s)}{\partial s} \tag{9}$$

We assume 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 follows:

$$\tilde{\chi}(s,t) \approx f(s+(1+\zeta)\delta(t)) \tag{10}$$

It shows that the processing magnifies motions, *i.e.* the spatial displacement $\delta(t)$ of the local feature $f(s)$ at time t, is amplified to a magnitude of $(1+\zeta)$.

$\delta(t)$ is not entirely within the passband of the temporal filter all the time. In this case, let $\delta_k(t)$, indexed by k, denote different

temporal spectral components of $\delta(t)$. Each $\delta_k(t)$ is attenuated by the temporal filtering by a factor γ_k. It produces a bandpass signal,

$$\phi(s,t) = \sum_k \gamma_k \delta_k(t) \frac{\partial f(s)}{\partial s} \tag{11}$$

Temporal frequency dependent attenuation can be equivalently interpreted as a frequency-dependent motion magnification factor, $\zeta_k = \gamma_k \zeta$, due to the multiplication in Eqn. (6), and the amplified output is

$$\tilde{\chi}(s,t) \approx f(s + \sum_k (1+\zeta_k)\delta_k(t)) \tag{12}$$

This procedure is summarized in Alg.2 and Fig.3(b) displays the examples of facial deformation trajectories before (green) and after (red) magnification.

5 Auto Key-frame Detection in 3D Videos

As we know, the facial expression is a quasi periodical signal, and another key issue in 4D *FER* is to segment the frames in the video from its occurrence to vanishing, discarding irrelevant ones which may incur disturbances. This is necessary to deal with long videos captured in real applications that include multiple expressions.

Generally, an expression consists of three main parts, *i.e.* onset, apex, and offset, and the frames within these parts are called key frames. As it can be seen from the left picture in Fig.5, all the six facial expressions of an individual randomly selected share such a similar process. In Fig.5, the deformation of a frame (*i.e. DSF*) is measured by its geodesic distance to the reference frame, related to certain expressions. We can also observe that the deformation intensity depends on the realization and the expression itself. For example, *SU* reflects the highest amount of 3D deformations since the mouth is often largely opened (as in Fig.6), while the lowest deformation is given by *SA* compared to the others.

In this section, a manifold clustering based key frame detection method is proposed to localize the onset, apex, and offset part of expressions in 3D videos. It is inspired by the consistency method [41], [42], a semi-supervised learning technique. For a given *DSF* trajectory $\|\chi(i)\| \in \mathcal{R}^m$ with $i = \{1, \cdots, n\}$, $\mathcal{L} = \{1, \cdots, c\}$ is a label set, and each point (*i.e. DSF*) only possesses one label. In a semi-supervised learning framework, the first l points $(1 \cdots l)$ are labeled and the other points $(l+1 \cdots n)$ are unlabeled. Then, we define $Y \in \mathcal{N}^{n \times c}$ with $Y_{ij} = 1$ if $\|\chi(i)\|$ has label j and 0 otherwise. n and c is the sample and cluster number, respectively. Let $\mathcal{F} \subset \mathcal{R}^{n \times c}$ denote all the matrices with non-negative entries. $F = [F_1^T, \cdots, F_n^T] \in \mathcal{F}$ is a matrix that labels all points $\|\chi(i)\|$ with a label $y_i = \text{argmax}_{j \leq c} F_{ij}$. We further define the series $F(t+1) = \varrho S F(t) + (1-\varrho)Y$ with $F(0) = Y$, $\varrho \in (0,1)$. The similarity matrix S can be represented as Eqn. (13),

$$S = D^{-1/2} W D^{-1/2} \tag{13}$$

where

$$D_{ij} = \begin{cases} \Sigma_{j=1}^n W_{ij}, i = j \\ 0, i \neq j \end{cases} \tag{14}$$

$$W_{ij} = \begin{cases} e^{\frac{-\|\|\chi(i)\| - \|\chi(j)\|\|^2}{2\sigma^2}}, i \neq j \\ 0, i = j \end{cases} \tag{15}$$

Let F^* be the limit of sequence $F(t)$. We label each point $\|\chi(i)\|$ as $y_i = \text{argmax}_{j \leq c} F_{ij}^*$, and further reach $F^* = (1-\varrho)(I - \varrho S)^{-1} Y$. [41] gives more details on this regularization framework and the closed form expression F^*.

We assume that σ, ϱ, c are known and that each cluster exposes a manifold without holes, *i.e.*, assigning one labeled point per class for the consistency method allows to find all the remaining points of individual classes. According to F^* computation, the solution to the semi-supervised learning problem only relies on the labels after $(I - \varrho S)$ is inverted. To adapt it to key frame detection, we transform the consistency method into a clustering algorithm. In this case, we find the centroid point that is the center of each class, and then determine whether other points belong to the class using $y_i = \text{argmax}_{j \leq c} F_{ij}$. We define a matrix U as:

$$U = (1-\varrho)(I - \varrho S)^{-1} = [u_1^T, \cdots, u_n^T] \tag{16}$$

where U defines a graph or diffusion kernel as described in [43]. The values of u_i^T in U are used to rank with respect to the intrinsic manifold structure as in [41]. The ordering of these distances along each manifold is maintained to be independent to scaling. Hence, without loss of ranking, the normalized form of U, called V, can be represented as:

$$V = [\frac{u_1^T}{\|u_1^T\|}, \cdots, \frac{u_n^T}{\|u_n^T\|}] = [v_1^T, \cdots, v_n^T] \tag{17}$$

V allows to define a rank based similarity metric between any points $\|\chi(i)\|$ and $\|\chi(j)\|$. For notational convenience, we further define the matrix D_M as:

$$D_M = \begin{bmatrix} 1 - v_1 v_1^T & \dots & 1 - v_1 v_n^T \\ \vdots & \ddots & \vdots \\ 1 - v_n v_1^T & \dots & 1 - v_n v_n^T \end{bmatrix} = [D_{M_1}, \dots, D_{M_n}] \tag{18}$$

In clustering, we pick clusters of points that are most similar to one another and most different from points in other clusters. We thus calculate the average of the columns of D_M and define an outlier vector O_d as follows, where O_{d_i} is the average distance between $\|\chi(i)\|$ and all the other points.

$$O_d = \left[\frac{1}{n}\sum D_{M_1}, \cdots, \frac{1}{n}\sum D_{M_n}\right] = [O_{d_1}, \cdots, O_{d_n}] \tag{19}$$

Alg.3 shows the procedure to identify the centroid points used to assign all the data points to a class. Let $\|\chi(l_1)\|, \cdots, \|\chi(l_c)\|$ be the centroid points of individual clusters. $\|\chi(l_1)\|$ is assigned to the point that is closest to all other points, which has the largest value O_{d_i}. To find $\|\chi(l_2)\|$, we multiply each element of O_d by the corresponding element in $D_{M_{l_1}}^T$, to obtain a new, re-weighted vector O_d. Denote O_d^n the n-th re-weighting of O_d. Re-weighting the vector gives all the points that are similar a small value and all the points that are different a large value. The point with the largest value O_d^n is selected, and the procedure of re-weighting and finding the most similar points repeats until c points are found.

For other points to be clustered, taking $\|\chi(k)\|$ as an example, we calculate W_{kj}, D_{kj} and S_{kj} ($j = \{1, \dots, n\}$), and produce W_k, D_k and S_k. We normalize the rows of S to the length of 1, denoted as S^1, and obtain the coefficients $\Theta = (\vartheta_1, \cdots, \vartheta_n)^T = S^1 (S_k^1)^T$. The vector has the property that if $\|\chi(k)\| = \|\chi(i)\|$, $\vartheta_i = 1$, but if $\|\chi(k)\|$ is far away from $\|\chi(i)\|$, ϑ_i will approach

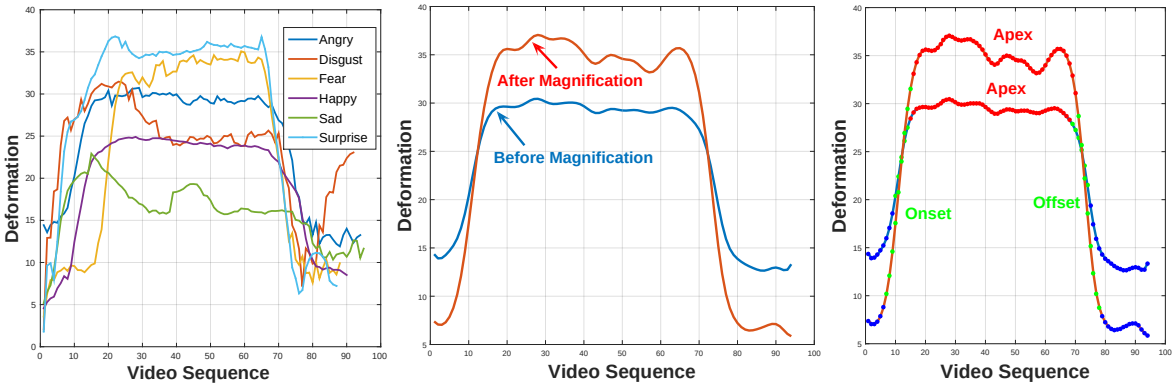


Fig. 5: The first column gives some examples of facial expression deformations in the BU-4DFE database. The second column shows the deformations before and after magnification of an angry sequence. The third column shows the clustering results on deformation manifolds, and the key-frames consist of the *Onset/Offset*-frames and *Apex*-frames.

Algorithm 3: Find Cluster Center Points

Input: Matrix D_M, number of clusters c

1: $n \leftarrow 1$, compute $O_d^1 = O_d$

2: $l_1 \leftarrow$ index of the example with maximum of O_d^1.

3: $w \leftarrow 1 - D_{M_{l_1}}$

4: **for** $2 \leq n \leq c$ **do**

  $l_n \leftarrow$ index of the example with maximum of O_d^n.

  $w \leftarrow (1 - D_{M_{l_n}}) \,.*\, w$

  $O_d^{n+1} \leftarrow w \,.*\, O_d^n$

Output: Indices of the clusters $l_1, \cdots, l_c$

zero. Therefore, ϑ_i measures the closeness of $\|\chi(k)\|$ to $\|\chi(i)\|$ in S. We use this property to assign $\|\chi(k)\|$ to a cluster by creating $F_k = [v_{l_1}\Theta^T, \cdots, v_{l_c}\Theta^T]$, where v_{l_i} denote the columns of V, corresponding to the cluster seed points. $\|\chi(k)\|$ is then assigned to a cluster $y_c = \text{argmax}_{j \leq c} F_k$, where j refers to an element in F_k. Fig.5 shows an example of clustering on the deformation manifolds with an angry expression and these detected key-frames make up the onset, apex, and offset parts. Fig.7 visualizes the key frames.

6 Experimental Results

This section describes the experiments conducted to validate the proposed method for 4D *FER*. A brief description of the BU-4DFE and BP4D datasets is first presented followed by the experimental setting and the classification accuracy, when including different steps of the pipeline depicted in Fig.1. A comparative study with existing methods is also displayed.

6.1 Dataset Description and Experimental Settings

Two benchmarks are adopted to evaluating the proposed approach in 4D *FER*, *i.e.* BU-4DFE and BP4D. The former focuses on a posed scenario and is used as the major database in performance analysis, while the latter considers a spontaneous scenario and is adopted to test the generalization ability. The databases as well as protocols are described in the subsequent.

BU-4DFE [32] is a 3D facial expression video dataset widely used for 3D and 4D *FER*. It contains in total 606 3D videos of 101 subjects, among which 58 are females and 43 are males, with 6 universal facial expressions for an individual. Each 3D sequence captures a facial expression at a rate of 25 fps (frames per second) and lasts approximately 3 to 4 seconds. Expressions are performed gradually including the temporal intervals *Onset-Apex-Offset*. 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 tip of the nose; a total of 50 sampled vertices on each curve is considered. Based on this parameterization step, the 3D face shapes along the video sequence are compared to a neutral reference frame (f^0) to derive the *DSF* feature, $\chi(t)$. Then, in spatial processing, Gaussian pyramid decomposition is used to decompose χ into 4 band levels. Finally, temporal processing to all the bands is applied, with the factor ζ set at 10, the sample rate ξ at 25, the attenuation rate γ at 1, and $\mathfrak{f} \in [0.3, 0.4]$. The frames on the expression trajectory manifold are clustered automatically according to the deformation $\|\chi(t)\|$, and the cluster number c is set to 3 (*Neutral*, *Onset/Offset*, and *Apex*). Our experiments are conducted on the following sub-pipelines: (1) the whole video sequence (denoted by *WV*), (2) the magnified whole video sequence (denoted as *MWV*), (3) the key frames of the video sequence (denoted by *KFV*), and finally (4) the magnified key-frames of the video sequence (*i.e. MKFV*).

As in [14], $\bar{\chi} = \frac{1}{n}\sum_{t=1}^{n}\chi(t)$ measures the overall expression deformation of the facial surfaces, where n denotes video length. Note a major difference to [14] where *DSFs* are computed between successive frames, they are quantified between the current frame and the reference in our approach. In addition, to concentrate more on the magnification step, instead of using a *Random Forest* (*RF*) classifier [14], *Support Vector Machine* (*SVM*) and *Hidden Markov Model* (*HMM*), more widely used for such an issue, are considered here, where the average *DSF* ($\bar{\chi}$) and the original *DSFs* ($\chi(t)$) are taken as the final feature vectors for expression label prediction, respectively. To achieve fair comparison with the previous studies, we randomly select 60 subjects (25 males and 35 females) from the BU-4DFE dataset to perform our experiments under a 10-fold cross-validation protocol.

BP4D is another dataset in this domain and is composed of 3D face videos belonging to 41 individuals, who are asked to perform 8 tasks, corresponding to 8 expressions. Besides the 6 universal ones, there are two additional expressions, namely embarrassment and pain. The data in BP4D are very similar to those in BU-4DFE; however, the facial expressions are elicited by a series of activities, including film watching, interviews, experiencing a cold pressor test, *etc.*, and are thus spontaneous. The experiment is carried out in a cross-dataset way, *i.e.*, training on BU-4DFE and testing on

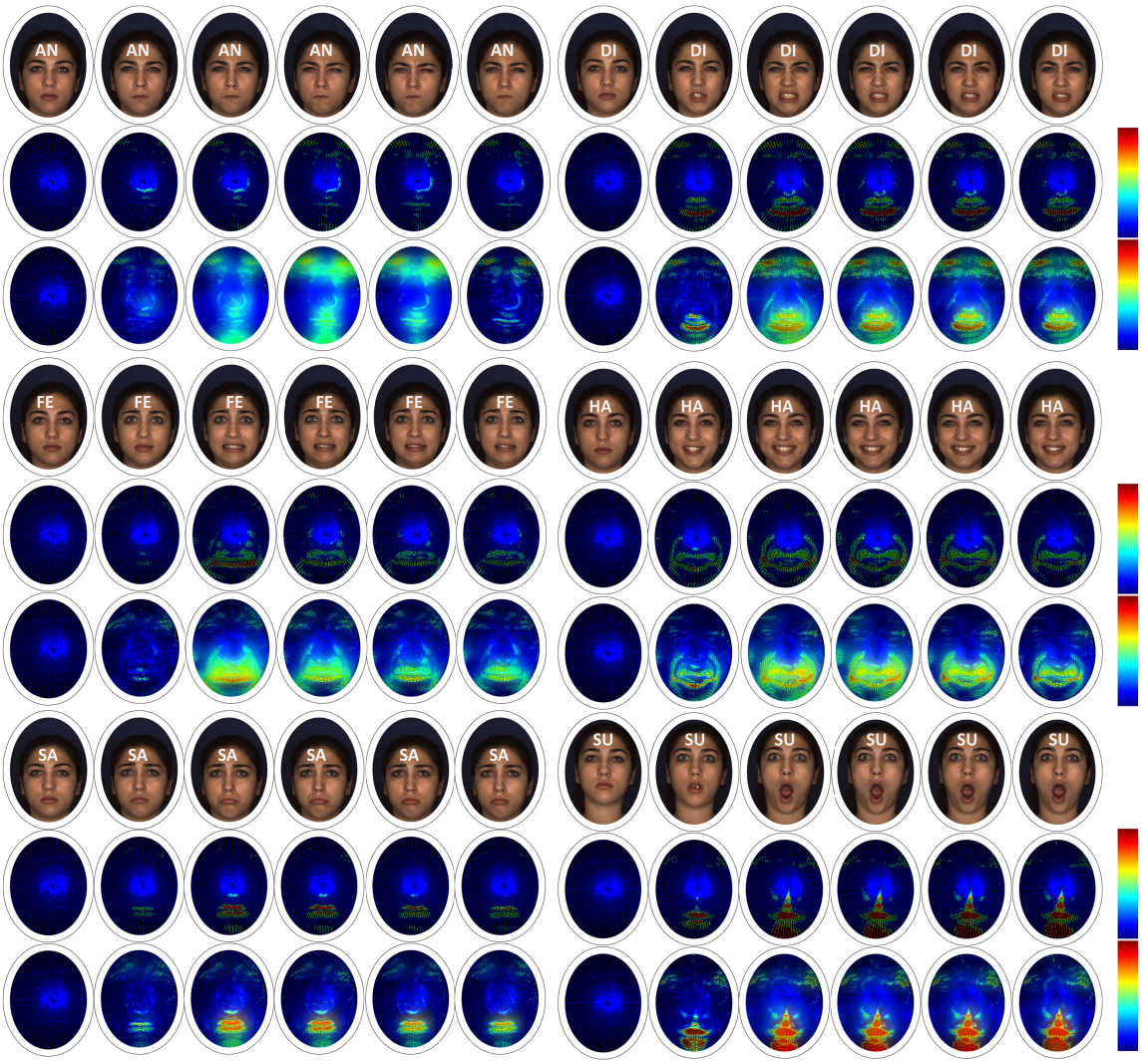


Fig. 6: Illustration of deformation magnification on sequences of the same subject performing the six universal expressions. One can appreciate the magnification effects on 3D deformations (third row) compared to the original facial deformation features (*DSFs*) (second row). Texture images corresponding to 3D face models are also displayed.

BP4D, according to the protocol used in [44]. The parameters are the same as the ones in the previous experiments on BU-4DFE.

6.2 Performance on BU-4DFE

We present the classification performance of *SVM on* $\bar{\chi}$ and *HMM on* $\chi(t)$ under different configurations, *i.e. WV*, *MWV*, *KFV*, and *MKFV*.

Table 1 provides a summary of our results. The results demonstrate the interest of the two contributions introduced in this study. Firstly, using the magnification step, an improvement that exceeds 10% in classification performance is achieved. Secondly, considering only the temporal interval of *Onset-Apex-Offset* automatically located reaches comparable results than utilizing full videos. This allows a large decrease in time consuming when performing 4D *FER* as it limits processing to only expression related parts instead of the entire sequence. For in-depth discussion of our results, we report in Table 2 and Table 3 the confusion matrices obtained for each of our schemes.

Before magnification, the proposed approach achieves 82.49% and 83.19% using *SVM* and *HMM* respectively, when full videos are input. Similar results are reached using the detected key video frames, *i.e.* 81.90% and 82.67%, respectively. Specifically, the *SU* and *HA* sequences are better classified with higher classification rates. This is mainly due to the high intensities and clear patterns of the deformations. The remaining expressions (*DI*, *FE*, *AN*, and *SA*) are harder to distinguish. We believe that two major reasons induce this difficulty: (1) the intra-class variability which confuses similar classes such as *DI*/*AN*/*FE*; (2) the low deformation magnitude exhibited when performing these expressions.

After magnification, the overall accuracies increase averagely more than 10% in all the experimental settings as Table 1 shows, which highlights the necessity to reveal subtle facial deformations. It can also be seen from these confusion matrices in Table 2 and Table 3, a significant performance gain is obtained in distinguishing *AN*, *DI*, and *FE*, the expressions that are reputed to be difficult to recognize. This quantitatively demonstrates the capability of the magnification step in 4D *FER*. These facts confirm the qualitative

TABLE 1: Average accuracies and standard deviations achieved by *SVM* and *HMM* on whole video sequences, only *apex* frames, and *onset-apex-offset* frames before and after magnification.

Algorithm	**Magnification**	**Whole Sequence** (%)	**Apex** (%)	**Onset-Apex-Offset** (%)
SVM on $\bar{\chi}$	N	82.49 ± 3.10	80.20 ± 2.37	81.90 ± 2.90
	Y	93.39 ± 3.54	89.50 ± 3.02	$\mathbf{94.46 \pm 3.34}$
HMM on $\chi(t)$	N	83.19 ± 2.83	81.10 ± 3.05	82.67 ± 2.91
	Y	94.18 ± 2.46	90.62 ± 2.93	$\mathbf{95.13 \pm 2.54}$

TABLE 2: The confusion matrices (*WV*, *MWV*, *KFV*, and *MKFV*) achieved by SVM on $\bar{\chi}$.

	Whole Video (*WV*)						**Key-Frames Video (*KFV*)**					
%	AN	DI	FE	HA	SA	SU	AN	DI	FE	HA	SA	SU
AN	**73.86**	9.18	6.49	1.85	6.11	2.51	**72.08**	9.36	6.43	1.86	6.12	4.15
DI	8.77	**71.27**	9.40	3.51	4.84	2.21	9.19	**69.55**	9.76	3.93	5.14	2.43
FE	5.89	5.37	**73.14**	4.59	5.39	5.62	5.78	5.37	**72.08**	4.99	5.60	6.18
HA	0.82	1.19	2.43	**93.60**	1.08	0.88	0.73	1.12	2.34	**94.16**	1.20	0.45
SA	2.55	2.28	2.99	1.65	**88.75**	1.78	2.39	2.14	2.84	1.56	**89.40**	1.67
SU	0.74	0.88	1.91	0.76	1.39	**94.32**	0.76	0.91	1.75	1.02	1.43	**94.13**
Average	**82.49 ± 3.10**						**81.90 ± 2.90**					
	Magnified Whole Video (*MWV*)						**Magnified Key-Frames Video (*MKFV*)**					
%	AN	DI	FE	HA	SA	SU	AN	DI	FE	HA	SA	SU
AN	**91.07**	2.73	2.01	1.59	2.08	0.52	**90.85**	3.12	2.08	0.92	2.14	0.89
DI	2.05	**92.62**	2.63	1.07	1.39	0.24	1.35	**94.58**	1.75	0.96	0.91	0.45
FE	1.66	1.53	**92.33**	1.32	1.54	1.62	1.24	1.16	**94.27**	0.88	1.58	0.87
HA	0.91	0.88	2.37	**94.29**	0.97	0.58	0.41	0.73	1.51	**96.34**	0.62	0.39
SA	1.37	1.22	1.62	0.90	**93.93**	0.96	1.21	1.09	1.43	0.78	**94.63**	0.86
SU	0.51	0.61	1.29	0.52	0.96	**96.11**	0.51	0.76	1.18	0.51	0.95	**96.09**
Average	**93.39 ± 3.54**						**94.46 ± 3.34**					

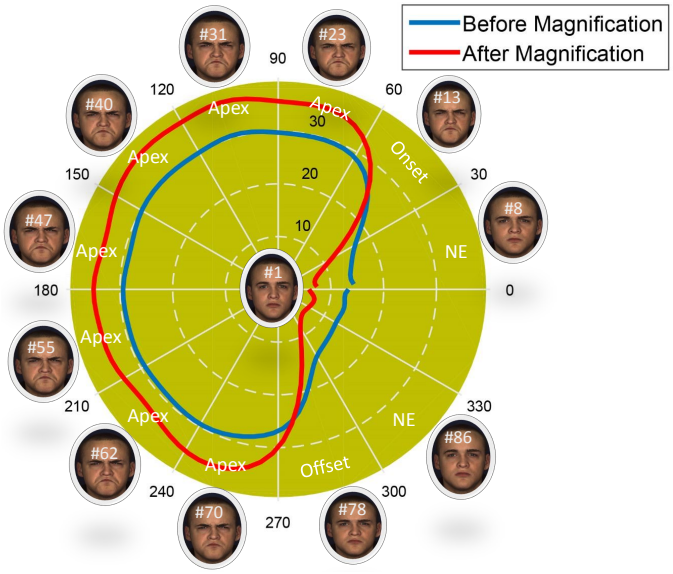


Fig. 7: Visualization of key frames detected on an anger sequence given in Fig.5.

results illustrated in Fig.6 where the deformations are amplified compared to the original features. The color-maps obtained after magnification reflect high amplitude of deformed areas which are not visible in original features.

6.3 Quantitative Analysis on Key Frames

To the best of our knowledge, the majority of current investigations in 4D *FER* consider either a full sequence or a short clip cropped through a sliding window of a pre-defined length as the input. The former methods generally assume that each sequence only records a single facial expression as the case in BU-4DFE, which tends to be problematic in real applications. The latter methods sample a part of the sequence at a time, which usually contains the frames of the neutral expression and thus has negative impacts on the final performance. Moreover, they also require a proper window size to optimize the precision. The proposed method addresses this issue by segmenting the key frames, *i.e.*, the *Onset-Apex-Offset* temporal interval, most relevant to the facial expression. When processing long sequences with multiple expressions, it is useful to locate this interval for more efficient analysis.

To evaluate the quality of the key-frame detector, we measure how the *Onset-Apex-Offset* parts automatically located are consistent to the ones manually labeled. Specifically, we randomly select 5 females and 5 males from the 41 subjects who are not involved in the experiments. 10 participants are asked to manually label the beginnings of the onsets and the endings of the offsets of the 60 sequences. In order to reduce subjective impacts, the average of all corresponding manual labels of each video is regarded as ground truth.

In total, there are 4229 frames in the ground truth *Onset-Apex-Offset* parts (5919 frames for all sequences). Before magnification, 4068 frames are automatically detected, where all are overlapped with ground truths. The precision and recall are 100% and 96.19% respectively. After magnification, 4318 frames are detected, where 4179 are overlapped. The precision and recall are 96.78% and

TABLE 3: The confusion matrices (*WV*, *MWV*, *KFV*, and *MKFV*) achieved by *HMM* on $\chi(t)$.

	Whole Video (*WV*)						**Key-Frames Video (*KFV*)**					
%	AN	DI	FE	HA	SA	SU	AN	DI	FE	HA	SA	SU
AN	**75.29**	5.88	7.31	1.14	8.17	2.21	**72.98**	6.43	8.12	1.13	8.93	2.41
DI	10.42	**71.55**	11.43	1.82	4.27	0.51	10.64	**70.97**	11.67	1.86	4.35	0.51
FE	5.07	6.86	**73.69**	3.33	8.06	2.99	5.28	7.15	**72.59**	3.47	8.39	3.12
HA	0.48	0.87	1.54	**94.93**	1.81	0.37	0.54	0.98	1.72	**94.33**	2.02	0.41
SA	3.71	1.01	4.18	0.65	**89.19**	1.26	3.26	0.95	3.90	0.73	**89.89**	1.27
SU	0.49	0.34	2.79	0.32	1.59	**94.47**	0.35	0.28	2.45	0.27	1.36	**95.29**
Average	**83.19 ± 2.83**						**82.67 ± 2.91**					
	Magnified Whole Video (*MWV*)						**Magnified Key-Frames Video (*MKFV*)**					
%	AN	DI	FE	HA	SA	SU	AN	DI	FE	HA	SA	SU
AN	**91.87**	1.92	2.41	0.38	2.69	0.73	**91.48**	2.03	2.51	0.39	2.83	0.76
DI	2.11	**94.22**	2.32	0.29	0.87	0.19	1.57	**95.74**	1.71	0.28	0.63	0.07
FE	1.38	1.86	**92.85**	0.91	2.19	0.81	1.08	1.46	**94.39**	0.72	1.71	0.64
HA	0.47	0.77	1.44	**95.30**	1.67	0.35	0.22	0.42	0.74	**97.59**	0.86	0.17
SA	1.85	0.51	2.07	0.33	**94.61**	0.63	1.60	0.43	1.79	0.28	**95.36**	0.54
SU	0.33	0.23	1.89	0.22	1.08	**96.25**	0.33	0.22	1.91	0.22	1.08	**96.24**
Average	**94.18 ± 2.46**						**95.13 ± 2.54**					

98.82% respectively. This answers the question that are probably raised when analyzing the results in Table 1 that the performance based on original *DSF* features is not as good as that using entire sequences. The *Onset-Apex-Offset* parts detected based on original *DSFs* are not complete enough with a relatively low recall, which indicates some important information is lost in facial expression representation. When launching the detector on magnified *DSFs*, the recall increases, preserving almost all the clues for *FER*. At the same time, the detector helps to discard most of the frames that are not related to expressions but bring in interference in decision, leading to a performance gain.

Another important conclusion made from Table 1 lies in that the *Onset/Offset* temporal interval contributes to classification. As we can see, taking the magnified features and the *SVM* classifier, the accurate based on the *Apex* frames (only) is 89.5% while this rate passes to 94.46% when considering the *Onset/Offset* frames. A similar case appears in *HMM* based prediction. It demonstrates the relevance of such facial dynamics in *FER*.

6.4 Comparison with State-of-the-art

In literature, a number of studies report 4D *FER* results on the BU-4DFE database; however they differ in their experimental settings. In this section, we compare our method to the existing ones when considering these differences.

The state of the art results reported on the BU-4DFE dataset are demonstrated in Table 4. In this table, #E means the number of expressions, #S is the number of subjects, #-CV denotes the fold number of cross-validation, and *Full Seq./Win* indicates the decision is made based on the analysis of full sequences or sub-sequences captured by a sliding window.

Sun *et al.* [12], [32] reach very competitive accuracies (90.44% and 94.37%) by using a sliding window of 6 frames; nevertheless, their approach requires manual annotation of 83 landmarks on the first frame. Furthermore, the vertex-level dense tracking technique is time consuming. In a more recent work from the same group developed by Reale *et al.* [37], the authors propose a 4D (*Space-Time*) feature termed *Nebula* calculated on a fixed-size window of 15 frames. The classification accuracy is 76.1% by using *SVM* on the onset part of each expression manually segmented. Sandbach *et al.* employ *3D-FFDs* for 3D facial shape motion representation and *HMMs* for classification. Variable window sizes are used for different expressions and the best score is 64.4%, achieved based on manual width values. Regarding the full sequence based ones, Le *et al.* [45] evaluate their algorithm which makes use of level curves and *HMMs* for classification in terms of three expressions (*HA*, *SA*, and *SU*) and display an accuracy of 92.22%. In [10], Fang *et al.* introduce *AFM* to register facial surfaces in the video sequence and *LBP-TOP* is then utilized to encode shape changes across time axis, reporting the performance of 74.63%. Xue *et al.* [36] exploit *3D-DCT* extracted from local depth patch-sequences according to automatically detected facial landmarks as spatio-temporal features, and the accuracy of 78.8% is achieved in *HMM* based prediction. Berretti *et al.* [35] compute local quantities and mutual distances from a set of auto-localized points and feed them into *HMMs*, reaching the result of 79.4%. Among the counterparts in Table 4, except [10] and [37] that carry out experiments on 100 subjects, the others randomly select 60 for evaluation.

TABLE 4: A comparative study of the proposed approach with the state-of-the-art on BU-4DFE.

Method	**Experimental Settings**	**Accuracy**
Sun and Yin [32]	6E, 60S, 10-CV, Win=6	90.44%
Sun *et al.* [12]	6E, 60S, 10-CV, Win=6	94.37%
Reale *et al.* [37]	6E, 100S, –, Win=15	76.10%
Sandbach *et al.* [3]	6E, 60S, 6-CV, Variable Win	64.60%
Fang *et al.* [10]	6E, 100S, 10-CV, Full Seq.	74.63%
Le *et al.* [45]	3E, 60S, 10-CV, Full Seq.	92.22%
Xue *et al.* [36]	6E, 60S, 10-CV, Full Seq.	78.80%
Berretti *et al.* [35]	6E, 60S, 10-CV, Full Seq.	79.44%
Berretti *et al.* [35]	6E, 60S, 10-CV, Win=6	72.25%
Ben Amor *et al.* [14]	6E, 60S, 10-CV, Win=6	93.83%
This work – *SVM* on $\bar{\chi}$	6E, 60S, 10-CV, Key Frames	**94.46%**
This work – *HMM* on $\chi(t)$	6E, 60S, 10-CV, Key Frames	**95.13%**

Compared to most of the methods listed in Table 4 under the protocol of 60 subjects, the proposed one obtains better results, up to 94.46% and 95.13% using *SVM* and *HMM* respectively. Recall that our main contribution lies in temporal magnification for geometry feature, which allows revealing subtle shape deformations

and thus delivering better discriminative power in distinguishing expressions. In addition, the introduction of automatic key-frame detection (*onset-apex-offset* frames) avoids exhaustive analysis by locating the most relevant parts in the 3D sequence.

6.5 Cross-dataset Evaluation on BP4D

To validate the generalization ability of the proposed method, we launch a cross-dataset experiment on BP4D according to [44]. BU-4DFE is employed for training and a subset of BP4D (*i.e.* Task 1 and Task 8, consisting of happy and disgust faces) for testing. It is actually a ternary classification problem on the samples of happy, disgust and neutral expressions. For computation simplicity, *SVM* is adopted as the classifier.

Table 5 demonstrates the results with and without the magnification step of the proposed method on the BP4D database, where similar conclusions can be drawn as in BU-4DFE. Compared with the performance in [44], the score on the original *DSF* features are better. When the deformations are further enhanced, the accuracy is improved to 81.7%, delivering a 10% result gain. It indicates that our method has the potential to be generalized to spontaneous situations.

7 Conclusions

This paper proposes a spatio-temporal processing method for 4D *FER*. It focuses on two important issues, *i.e.* expression related frame detection and subtle deformation magnification. After a pre-processing step, the flow of 3D faces is generated to capture spatial deformations based on the Riemannian theory, where registration and comparison are fulfilled at the same time. The deformations obtained are then amplified by using the temporal filter over the 3D face video. The combination of these two ideas performs accurate vertex-level registration of 4D faces and highlights hidden shape variations in 3D face video sequences. Furthermore, we present an automatic technique to localize relevant frames (*onset-apex-offset* frames) in the video based on clustering facial shape deformation manifolds. Through comprehensive experiments, we demonstrate the contribution of such joint spatio-temporal analysis in recognizing facial expressions from 4D data. The proposed approach outperforms the existing ones on the BU-4DFE dataset and shows good potential to be generalized to spontaneous scenarios as in BP4D.

In future work, we will continue studying the way to improve the performance of BP4D as well as more practical benchmarks.

References

[1] T. Fang, X. Zhao, O. Ocegueda, S. K. Shah, and I. A. Kakadiaris, "3d facial expression recognition: A perspective on promises and challenges," in *IEEE International Conference on Automatic Face and Gesture Recognition*, 2011, pp. 603–610.

[2] G. Sandbach, S. Zafeiriou, M. Pantic, and L. Yin, "Static and dynamic 3d facial expression recognition: A comprehensive survey," *Image Vision Computing*, vol. 30, no. 10, pp. 683–697, 2012.

[3] G. Sandbach, S. Zafeiriou, M. Pantic, and D. Rueckert, "Recognition of 3d facial expression dynamics," *Image and Vision Computing*, vol. 30, no. 10, pp. 762–773, 2012.

[4] H. Li, H. Ding, D. Huang, Y. Wang, X. Zhao, J.-M. Morvan, and L. Chen, "An efficient multimodal 2d + 3d feature-based approach to automatic facial expression recognition," *Computer Vision and Image Understanding*, vol. 140, pp. 83–92, 2015.

[5] P. Ekman and W. Friesen, *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Salt Lake City: Consulting Psychologists Press, 1978.

[6] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39–58, 2009.

[7] B. Fasel and J. Luettin, "Automatic facial expression analysis: a survey," *Pattern Recognition*, vol. 36, no. 1, pp. 259–275, 2003.

[8] W. Zheng, "Multi-view facial expression recognition based on group sparse reduced-rank regression," *IEEE Transactions on Affective Computing*, vol. 5, no. 1, pp. 71–85, 2014.

[9] A. Maalej, B. Ben Amor, M. Daoudi, A. Srivastava, and S. Berretti, "Shape analysis of local facial patches for 3D facial expression recognition," *Pattern Recognition*, vol. 44, no. 8, pp. 1581–1589, 2011.

[10] T. Fang, X. Zhao, O. Ocegueda, S. K. Shah, and I. A. Kakadiaris, "3d/4d facial expression analysis: An advanced annotated face model approach," *Image and Vision Computing*, vol. 30, no. 10, pp. 738–749, 2012.

[11] X. Yang, D. Huang, Y. Wang, and L. Chen, "Automatic 3d facial expression recognition using geometric scattering representation," in *IEEE International Conference on Automatic Face and Gesture Recognition and Workshops*, vol. 1, 2015, pp. 1–6.

[12] Y. Sun, X. Chen, M. Rosato, and L. Yin, "Tracking vertex flow and model adaptation for three-dimensional spatiotemporal face analysis," *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, vol. 40, no. 3, pp. 461–474, 2010.

[13] T. Fang, X. Zhao, S. K. Shah, and I. A. Kakadiaris, "4d facial expression recognition," in *IEEE International Conference on Computer Vision Workshops*, 2011, pp. 1594–1601.

[14] B. Ben Amor, H. Drira, S. Berretti, M. Daoudi, and A. Srivastava, "4-d facial expression recognition by learning geometric deformations," *IEEE Transactions on Cybernetics*, vol. 44, no. 12, pp. 2443–2457, 2014.

[15] A. Jan and H. Meng, "Automatic 3d facial expression recognition using geometric and textured feature fusion," in *IEEE International Conference on Automatic Face and Gesture Recognition and Workshops*, 2015, pp. 1–6.

[16] Q. Zhen, D. Huang, Y. Wang, and L. Chen, "Muscular movement model based automatic 3d facial expression recognition," in *International Conference On MultiMedia Modeling*, 2015, pp. 522–533.

[17] C. Liu, A. Torralba, W. T. Freeman, F. Durand, and E. H. Adelson, "Motion magnification," *ACM Transactions on Graphics*, vol. 24, no. 3, pp. 519–526, 2005.

[18] J. Wang, S. M. Drucker, M. Agrawala, and M. F. Cohen, "The cartoon animation filter," *ACM Transactions on Graphics*, vol. 25, no. 3, pp. 1169–1173, 2006.

[19] M.-Z. Poh, D. J. McDuff, and R. W. Picard, "Non-contact, automated cardiac pulse measurements using video imaging and blind source separation," *Optics Express*, vol. 18, no. 10, pp. 10 762–10 774, 2010.

[20] H.-Y. Wu, M. Rubinstein, E. Shih, J. Guttag, F. Durand, and W. Freeman, "Eulerian video magnification for revealing subtle changes in the world," *ACM Transactions on Graphics*, vol. 31, no. 4, pp. 1–8, 2012.

[21] Q. Zhen, D. Huang, Y. Wang, H. Drira, B. Ben Amor, and M. Daoudi, "Magnifying subtle facial motions for 4d expression recognition," in *IEEE/IAPR International Conference on Pattern Recognition*, 2016, pp. 1–6.

[22] I. Mpiperis, S. Malassiotis, and M. G. Strintzis, "Bilinear models for 3-d face and facial expression recognition," *IEEE Transactions On Information Forensics and Security*, vol. 3, no. 3, pp. 498–511, 2008.

[23] O. Ocegueda, T. Fang, S. K. Shah, and I. A. Kakadiaris, "Expressive maps for 3d facial expression recognition," in *IEEE International Conference on Computer Vision Workshops*, 2011, pp. 1270–1275.

[24] X. Zhao, D. Huang, E. Dellandrea, and L. Chen, "Automatic 3d facial expression recognition based on a bayesian belief net and a statistical facial feature model," in *International Conference on Pattern Recognition*, 2010, pp. 3724–3727.

[25] X. Li, T. Jia, and H. Zhang, "Expression-insensitive 3d face recognition using sparse representation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 2575–2582.

[26] H. Li, L. Chen, D. Huang, Y. Wang, and J. Morvan, "3d facial expression recognition via multiple kernel learning of multi-scale local normal patterns," in *International Conference on Pattern Recognition*, 2012, pp. 2577–2580.

[27] C. Samir, A. Srivastava, M. Daoudi, and E. Klassen, "An intrinsic framework for analysis of facial surfaces," *International Journal of Computer Vision*, vol. 82, no. 1, pp. 80–95, 2009.

[28] H. Drira, B. Ben Amor, A. Srivastava, M. Daoudi, and R. Slama, "3d face recognition under expressions, occlusions, and pose variations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 9, pp. 2270–2283, 2013.

TABLE 5: Cross-dataset evaluation on the BP4D database.

Method	Training in BU-4DFE	Testing in BP4D	Accuracy (%)
Zhang *et al.* [44]	Happy, Disgust, Neutral	Task 1 and Task 8	71.0
This work (Before Magnification)	Happy, Disgust, Neutral	Task 1 and Task 8	75.6
This work (After Magnification)	Happy, Disgust, Neutral	Task 1 and Task 8	**81.7**

[29] L. Ballihi, B. Ben Amor, M. Daoudi, A. Srivastava, and D. Aboutajdine, "Boosting 3-d-geometric features for efficient face recognition and gender classification," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 6, pp. 1766–1779, 2012.

[30] B. Xia, B. Ben Amor, M. Daoudi, and H. Drira, "Can 3d shape of the face reveal your age?" in *International Conference on Computer Vision Theory and Applications*, 2014, pp. 5–13.

[31] M. Devanne, H. Wannous, S. Berretti, P. Pala, M. Daoudi, and A. D. Bimbo, "3-d human action recognition by shape analysis of motion trajectories on riemannian manifold," *IEEE Transactions on Cybernetics*, vol. 45, no. 7, pp. 1340–1352, 2015.

[32] Y. Sun and L. Yin, "Facial expression recognition based on 3d dynamic range model sequences," in *Europeon Conference on Computer Vision*, 2008, pp. 58–71.

[33] S. Canavan, Y. Sun, X. Zhang, and L. Yin, "A dynamic curvature based approach for facial activity analysis in 3d space," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2012, pp. 14–19.

[34] G. Sandbach, S. Zafeiriou, M. Pantic, and D. Rueckert, "A dynamic approach to the recognition of 3d facial expressions and their temporal models," in *IEEE International Conference on Automatic Face and Gesture Recognition and Workshops*, 2011, pp. 406–413.

[35] S. Berretti, A. Del Bimbo, and P. Pala, "Automatic facial expression recognition in real-time from dynamic sequences of 3d face scans," *The Visual Computer*, vol. 29, no. 12, pp. 1333–1350, 2013.

[36] M. Xue, A. Mian, W. Liu, and L. Li, "Automatic 4d facial expression recognition using dct features," in *IEEE Winter Conference on Applications of Computer Vision*, pp. 199–206.

[37] M. Reale, X. Zhang, and L. Yin, "Nebula feature: A space-time feature for posed and spontaneous 4D facial behavior analysis." in *IEEE International Conference on Automatic Face and Gesture Recognition*, 2013, pp. 1–8.

[38] M. Hayat and M. Bennamoun, "An automatic framework for textured 3d video-based facial expression recognition," *IEEE Transactions on Affective Computing*, vol. 5, no. 3, pp. 301–313, 2014.

[39] H. Drira, B. Ben Amor, M. Daoudi, A. Srivastava, and S. Berretti, "3d dynamic expression recognition based on a novel deformation vector field and random forest," in *International Conference on Pattern Recognition*, 2012, pp. 1104–1107.

[40] M. Daoudi, H. Drira, B. Ben Amor, and S. Berretti, "A dynamic geometry-based approach for 4d facial expressions recognition," in *European Workshop on Visual Information Processing*, 2013, pp. 280–284.

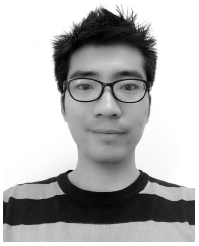
[41] D. Zhou, J. Weston, A. Gretton, O. Bousquet, and B. Schölkopf, "Ranking on data manifolds," in *Advances in Neural Information Processing Systems*, 2004, pp. 169–176.

[42] M. Breitenbach and G. Z. Grudic, "Clustering through ranking on manifolds," in *International Conference on Machine learning*, vol. 119, no. 8, 2005, pp. 73–80.

[43] J. Shrager, T. Hogg, and B. A. Huberman, "Observation of phase transitions in spreading activation networks," *Science*, vol. 236, no. 4805, pp. 1092–1094, 1987.

[44] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, and J. M. Girard, "Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database," *Image and Vision Computing*, vol. 32, no. 10, pp. 692–706, 2014.

[45] V. Le, H. Tang, and T. Huang, "Expression recognition from 3d dynamic faces using robust spatio-temporal shape features," in *IEEE International Conference on Automatic Face Gesture Recognition and Workshops*, 2011, pp. 414–421.



Qingkai Zhen received the B.S. in computer science from Hebei Institute Of Architecture and Civil Engineering, HeBei, China, in 2007, and M.S. degrees in computer science from Xiamen University, Xiamen, China, in 2011, and now a PH.D. candidate in the Laboratory of Intelligent Recognition and Image Processing with the Beijing Key Laboratory of Digital Media, School of Computer Science and Engineering, Beihang University, Beijing, China. His current research interests include biometrics, 2D/3D/4D face analysis, image/video processing, object tracking, and pattern recognition.

Di Huang (S'10-M'11) received the B.S. and M.S. degrees in computer science from Beihang University, Beijing, China, and the Ph.D. degree in computer science from the École centrale de Lyon, Lyon, France, in 2005, 2008, and 2011, respectively. He joined the Laboratory of Intelligent Recognition and Image Processing with the Beijing Key Laboratory of Digital Media, School of Computer Science and Engineering, Beihang University, as a Faculty Member. His current research interests include biometrics, in particular, on 2D/3D face analysis, image/video processing, and pattern recognition.

Yunhong Wang (M'98-SM'15) received the B.S. degree in electronic engineering from Northwestern Polytechnical University, Xian, China, in 1989, and the M.S. and Ph.D. degrees in electronic engineering from Nanjing University of Science and Technology, Nanjing, China, in 1995 and 1998, respectively. She was with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China, from 1998 to 2004. Since 2004, she has been a Professor with the School of Computer Science and Engineering, Beihang University, Beijing, where she is also the Director of Laboratory of Intelligent Recognition and Image Processing, Beijing Key Laboratory of Digital Media. Her current research interests include biometrics, pattern recognition, computer vision, data fusion, and image processing.

Hassen Drira received the PHD degree in computer science in 2011 from the University of Lille1, France. He is an assistant professor of computer science at Télécom Lille and has been a member of the Laboratoire CRIStAL (UMR CNRS 9189) since September 2012. His research interests are mainly focused on pattern recognition, statistical shape analysis. He has published several refereed journals and conference articles in these areas.

Boulbaba Ben Amor received the engineer degree in computer science from ENIS, Tunisia, in 2002, and the MS and PHD degrees in computer science both from the Ecole Centrale de Lyon, France, in 2003 and 2006, respectively. He jointed the Mines-Télécom/Télécom Lille1 Institute as an associate professor in 2007. Since then, he has also been a member of the Computer Science Laboratory at the University Lill 1 (LIFL UMR CNRS 8022). His research interests include statistical 3D face analysis and recognition and facial expression recognition using 3D. He is a coauthor of several papers in refereed journals and proceedings of international conference. He has been involved in French and International projects and has served as a program committee member and reviewer for international journals and conference. He is a member of the IEEE.

Mohamed Daoudi received the PHD degree in computer engineering from the University of Lille 1, France, in 1993 and the Habilitation à Diriger des Recherches from the University of Littoral, France, in 2000. He is a professor of computer science at TELECOM Lille 1 and LIFL (UMR CNRS 8022). He is the head of the Computer Science Department at Télécom Lille1. He was the founder and the scientific leader of the MIIRE research group http://www-rech.telecom-lille1.eu/miire/. His research interests include pattern recognition, image processing, 3D analysis and retrieval, and 3D face analysis and recognition. He has published more than 100 papers in some of the most distinguised scientific journals and international conferences. He is the coauthor of the book 3D Processing: Compression, Indexing and Watermarking (Wiley, 2008). He is a senior member of the IEEE.