3D Dynamic Expression Recognition Based on a Novel Deformation Vector Field and Random Forest

Drira Hassen, Boulbaba Ben Amor, Daoudi Mohamed, Srivastava Anuj, Stefano Berretti

3D Dynamic Expression Recognition based on a Novel Deformation Vector Field and Random Forest

Hassen Drira
LIFL (UMR Lille1/CNRS 8022), France.
hassen.drira@telecom-lille1.eu

Boulbaba Ben Amor, Mohamed Daoudi
LIFL (UMR Lille1/CNRS 8022), France.
Institut TELECOM ; TELECOM Lille 1, France.

Anuj Srivastava
Departement of Statistics, Florida State University, USA.

Stefano Berretti
Dipartimento di Sistemi e Informatica University of Firenze, Italy.

Abstract

This paper proposes a new method for facial motion extraction to represent, learn and recognize observed expressions, from 4D video sequences. The approach called ***Deformation Vector Field*** *(DVF) is based on Riemannian facial shape analysis and captures densely dynamic information from the entire face. The resulting temporal vector field is used to build the feature vector for expression recognition from 3D dynamic faces. By applying LDA-based feature space transformation for dimensionality reduction which is followed by a Multi-class Random Forest learning algorithm, the proposed approach achieved 93% average recognition rate on BU-4DFE database and outperforms state-of-art approaches.*

1 Introduction and related work

Facial expressions analysis and recognition from 2D and/or 3D data has emerged as an active research topic with applications in several different areas, such as human-centered user interfaces, computer graphics, facial animation, ambient intelligence, etc. In recent years, there has been tremendous interest in tracking and recognizing facial expressions over time, it is suggested that the dynamics of facial expressions provides important cues about the underlying emotions that are not available in static 2D or 3D images.

In 3D-based researches, there are a few works that use 4D data for facial expression analysis. In [7], a spatio-temporal expression analysis approach based on 3D dynamic geometric facial model sequences is proposed. The approach integrates a 3D facial surface descriptor and Hidden Markov Models (HMMs) to recognize facial expressions. Experiments were performed on the BU-4DFE and their approach achieved a 90.44% (using spatio-temporal) and 80.04% (using temporal only) recognition rates for distinguishing the six prototypic facial expressions (anger, disgust, fear, happiness, sadness and surprise). The main limitation of this solution resides in the use of 83 proprietary annotated landmarks of the BU-4DFE that are not released for public use. An extension of this work for 4D face recognition appeared in [6]. Sandbach et al. proposed in [3] to exploit the dynamics of 3D facial motions to recognize observed expressions. First, they captured motion between frames using Free-Form Deformations and extract motion features using a quad-tree decomposition of several motion fields. Second, they used HMM models for temporal modeling of the full expression sequence to be represented by 4 segments which are neutral-onset-apex-offset expression segments. In their approach, features selection is derived using GentleBoost technique applied on the onset and offset temporal segments of the expression, the average correct classification results for three basic expressions (happy, angry and surprise) achieved 81.93% in whole sequence-based classification and 73.61% in frame-by-frame classification. An extension of this work is presented in [4]. Another work on 4D facial expression recognition is proposed by Le et al. [2]. They proposed a level curve based approach to capture the shape of 3D facial models. The level curves are parameterized using the arc-length function. The Chamfer distances is applied to measure the distances between the corresponding normalized segments, partitioned from these level curves of two 3D facial shapes. These measures are then used as spatio-temporal features to train Hidden Markov Model (HMM), and since the training data were not sufficient for learning HMM, the authors propose to apply the universal background modeling

(UBM) to overcome the over-fitting problem. Using the BU-4DFE database to evaluate their approach, they reached an overall recognition accuracy of 92.22% for three prototypic expressions (happy, sad and surprise), when classifying whole sequences.

In the present work, we present a fully automatic facial expression recognition approach based on 4D (3D+t) data. To capture the deformation across the sequences, we propose a new method called *Deformation Vector Filed*, obtained via facial surfaces parametrization by collections of radial curves emanating from their tips of the noses. Our paper is organized as follows: In Sect. 2, a face representation model is proposed that captures facial features relevant to categorize expression variations in 3D dynamic sequences. In Sect. 3, the LDA-based feature space transformation and Multiclass Random Forest based classification are addressed. Experimental results and comparative evaluation obtained on the BU-4DFE are reported and discussed in Sect. 4. Finally, conclusions are outlined in Sect. 5.

2 The Deformation Vector Field

One basic idea to capture facial deformation across 3D video sequences is to track densely meshes' vertices along successive 3D frames. To do so, as the meshes resolutions vary across 3D video frames, establishing a dense matching on consecutive frames is necessary. Sun et al. [7] proposed to adapt a generic model (a tracking model) to each 3D frame. However, a set of 83 predefined key-points is required to control the adaptation based on radial basis function. A second solution is presented by Sandbach et al. [3], where the authors used an existing non-rigid registration algorithm (FFD) based on B-splines interpolation between a lattice of control points. The dense matching is a step of preprocessing stage to estimate a motion vector field between frames t and t-1. However, the results provided by the authors are limited to three facial expressions: *happy*, *angry* and *surprise*. To address this problem, we propose to represent facial surfaces by a set of parameterized radial curves emanating from the tip of the nose. Such an approximation of facial surfaces by indexed collection of curves can be seen as solution to facial surface parameterizations which capture locally their shapes.

2.1 Shape Deformation Capture

A parameterized curve on the face, $\beta : I \rightarrow \mathbb{R}^3$, where $I = [0,1]$, is represented mathematically using the *square-root velocity function* [5], denoted by $q(t)$, according to: $q(t) = \frac{\dot{\beta}(t)}{\sqrt{\|\dot{\beta}(t)\|}}$. This specific parametrization has the advantage of capturing the shape of the curve and provides simple calculus. Let define the space of such functions: $\mathcal{C} = \{q : I \rightarrow \mathbb{R}^3, \|q\| = 1\} \subset \mathbb{L}^2(I, \mathbb{R}^3)$, where $\|\cdot\|$ implies the $\mathbb{L}^2$ norm. With the $\mathbb{L}^2$ metric on its tangent spaces, $\mathcal{C}$ becomes a Riemannian manifold. Given two curves q_1 and q_2, let ψ denotes a path on the manifold $\mathcal{C}$ between q_1 and q_2, $\dot{\psi} \in T_\psi(\mathcal{C})$ is a tangent vector field on the curve $\psi \in \mathcal{C}$ and $< . >$ denotes the $\mathbb{L}^2$ inner product on the tangent space. In our case, as the elements of $\mathcal{C}$ have a unit $\mathbb{L}^2$ norm, $\mathcal{C}$ is a Hypersphere in the Hilbert space $\mathbb{L}^2(I, \mathbb{R}^3)$. The geodesic path between any two points $q_1, q_2 \in \mathcal{C}$ is simply given by the minor arc of great circle connecting them on this Hypersphere, $\psi^* : [0,1] \rightarrow \mathcal{C}$, given by Eq. (1):

$$\psi^*(\tau) = \frac{1}{\sin(\theta)} \left(\sin((1-\tau)\theta)q_1 + \sin(\theta\tau)q_2\right) \quad (1)$$

and $\theta = d_{\mathcal{C}}(q_1, q_2) = cos^{-1}(\langle q_1, q_2 \rangle)$. We point out that $sin(\theta) = 0$ if the distance between the two curves is null, in other words $q_1 = q_2$. In this case, for each τ, $\psi^*(\tau) = q_1 = q_2$. The tangent vector field on this geodesic $\frac{d\psi^*}{d\tau} : [0,1] \rightarrow T_\psi(\mathcal{C})$ is then given by Eq. (2):

$$\frac{d\psi^*}{d\tau} = \frac{-\theta}{\sin(\theta)} \left(\cos((1-\tau)\theta)q_1 - \cos(\theta\tau)q_2\right) \quad (2)$$

Knowing that on geodesic path, the covariant derivative of its tangent vector field is equal to 0, $\frac{d\psi^*}{d\tau}$ is parallel along the geodesic ψ^* and we shall represent it with $\frac{d\psi^*}{d\tau}|_{\tau=0}$. Accordingly, Eq. (2) becomes:

$$\frac{d\psi^*}{d\tau}|_{\tau=0} = \frac{\theta}{\sin(\theta)} \left(q_2 - \cos(\theta)q_1\right) \ , \quad (3)$$

with $\theta \neq 0$. Thus, $\frac{d\psi^*}{d\tau}|_{\tau=0}$ is sufficient to represent this vector field, the remaining vectors can be obtained by parallel transport of $\frac{d\psi^*}{d\tau}|_{\tau=0}$ along the geodesic ψ^*. In practice, the first step to capture the deformation between two given 3D faces S^1 and S^2 is to extract the radial curves. Let β^1_α and β^2_α denote the radial curves that make an angle α with a reference radial curve on faces S^1 and S^2, respectively. The reference curve is chosen to be the vertical curve as the faces have been rotated to the upright position during the preprocessing step. The tangent vector field $\dot{\psi_\alpha}^*$ that represents the energy E given in Eq. (1) needed to deform β^1_α to β^2_α is then calculated for each index α. We consider the magnitude of this vector field at each point, located in β_α and is of index k on this curve, for building a *Deformation Vector Field* on the facial surface, $V^k_\alpha = ||\dot{\psi^*_\alpha}|_{(\tau=0)}(k)||$, where α denotes the angle to the vertical radial curve and k denotes a point on this curve. This scalar vector field quantifies the local (on each point) deformation between the faces S^1 and S^2.

2.2 Dynamic Shape Deformation Analysis

To capture the dynamic of facial deformations across 3D face sequences, we consider the deformation map computed between successive frames, as described in section 2.1. Similarly to the recognition schema proposed by Sun et al. [7], in order to make possible to come to the recognition system at any time and make the recognition process possible from any frame of a given video, we consider subsequences of n frames. Thus, we chose the first n frames as the first subsequence. Then, we chose n-consecutive frames starting from the second frame as the second subsequence. The process is repeated by shifting the starting index of the sequence every one frame till the end of the sequence. For each sub-sequence, the first frame is considered as reference one and the deformation map is calculated to each of the remaining frames. The feature vector for this subsequence is built based on the average deformation of the $n-1$ calculated deformation maps. Thus, each subsequence is represented by a feature vector of size the number of points on the face. We point out that we consider 5000 points, 50 for each of the 100 radial curves.

Figure 1 illustrates one subsequence for each expression with $n = 6$ frames. Each expression is illustrated in two rows, the upper row gives the reference frame of the subsequence and the $n-1$ successive frames of subsequences. Below are the corresponding deformation maps computed for each frame. The mean deformation map is reported at the right and represent the feature vector for that subsequence. Thus, this deformation map summarizes the temporal deformation undergone by the facial surface when conveying expressions.

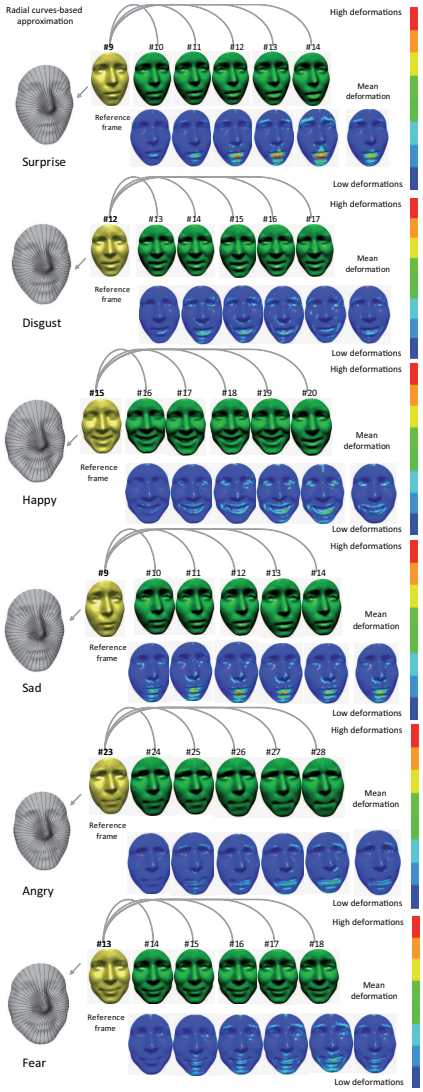

Figure 1. Calculus of dynamic shape deformation on subsequences taken from the BU-4DFE sequences.

3 Classification

We now represent each subsequence by its Deformation Vector Filed $V_\alpha^k = ||\dot{\psi}_\alpha^*|_{(\tau=0)}(k)||$, as described in section 2. Since, the dimensionality of the feature vector is high, we use LDA-based transformation (Linear Discriminant Analysis) to transform the present feature space to an optimal one that is relatively insensitive to different subjects while preserving the discriminative expression information. LDA defines the within-class matrix S_w and the between-class matrix S_b. It transforms a n-dimensional feature to an optimized d-dimensional feature where d < n. For our experiments, the discriminative classes are 6 expressions, thus the reduced dimension d is 5. For the classification task we used the Multi-class version of Random Forest algorithm. The Random Forest algorithm was proposed by Leo Breiman in [1] and defined as a meta-learner comprised of many individual trees. It was designed to operate quickly over large datasets and more importantly to be diverse by using random samples to build each tree in the forest. Diversity is obtained by randomly choosing attributes at each node of the tree and then using the attribute that provides the highest level of learning. Once trained, Random Forest classify a new expression from an input feature vector by putting it down each of the trees in the forest. Each tree gives a classification decision by voting for that class. Then, the forest chooses the classification having the most votes (over all the trees in the forest). In our experiments we used Weka Multi-class implementation of Random Forest al-

gorithm by considering 40 trees.

4 Experiments and Discussions

To demonstrate the effectiveness of the proposed approach, we perform expression recognition from 3D face sequences following the experimental setting described in [7]. The experiments were conducted on a subset of 60 subjects (arbitrarily selected) from the BU-4DFE dataset and includes the 6 prototypic expressions (Angry (An), Disgust (Di), Fear (Fe), Happy (Ha), Sad (Sa), Surprise (Su)). We note that the experiments are conducted in a identity-independent fashion. Following a similar setup as in [7], we randomly divided the feature vectors (computed from the subsequences) of 60 subjects into two sets, the training set containing 54 subjects, and the test set containing 6 subjects. We perform on these feature vectors Random Forest algorithm and reported obtained results in Table 1. We note that the reported rates are obtained by averaging the results of the 10-independent and arbitrarily run experiments (10-fold cross validation). The average recognition rate is equal to 93.21%. It can also be observed that the best classified expressions are (Ha) and (Su) with recognition accuracies of 95.47% and 94.53%, respectively, whereas the (Fe) expression is more difficult to classify. This is mainly due to the subtle changes in facial shapes motion of this expression compared to (Ha) and (Su).

Table 1. Confusion matrix.

%	*An*	*Di*	*Fe*	*Ha*	*Sa*	*Su*
An	**93.11**	2.42	1.71	0.46	1.61	0.66
Di	2.3	**92.46**	2.44	0.92	1.27	0.58
Fe	1.89	1.75	**91.24**	1.5	1.76	1.83
Ha	0.57	0.84	1.71	**95.47**	0.77	0.62
Sa	1.7	1.52	2.01	1.09	**92.46**	1.19
Su	0.71	0.85	1.84	0.72	1.33	**94.53**
Average recognition rate = 93.21%						

We note that the proposed approach outperforms state-of-the-art approaches following the same experimental settings. The recognition rates reported in [7] based on temporal analysis only was 80.04% and spatio-temporal analysis was 90.44%. In both studies subsequences of constant window width equal to 6 $(Win = 6)$ is defined for experiments. We emphasize that their approach is not completely automatic requiring 83 manually marked key points on the first frame of the sequence to allow accurate model tracking. We can also make comparisons with performances showed in [3] on only three expressions (*Ha*, *An*, and *Su*), their approach achieved an average recognition rate about 73.61% (for each frame) compared to our approach which showed an increase of more than 19% by achieving (93.21%). We note also that the subjects considered on our study are arbitrary selected whereas in [3] sequences are accurately selected. Approaches like [2] and [3] reported recognition results on whole facial sequences, this hinders the possibility of the methods to adhere to a real-time protocol. In fact, showing recognition results depends on the preprocessing of whole sequences unlike our approach and the one described in [7] which are able to provide recognition results when processing very few 3D frames.

5 Conclusion

This paper proposes a new Deformation Vector Field (DVF) which accurately describes local deformations across 3D facial sequences. A facial surface parametrization by their radial curves allows the definition of this descriptor on each facial point based on Riemannian Geometry. Then the well known learning algorithm, Random Forest, is performed for the classification task. Experiments conducted on BU-4DFE dataset, following state-of-the-art setting demonstrate the effectiveness of the proposed approach based only on temporal analysis of facial sequences.

References

[1] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[2] V. Le, H. Tang, and T. Huang. Expression recognition from 3d dynamic faces using robust spatio-temporal shape features. In *Automatic Face Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 414 –421, 2011.

[3] G. Sandbach, S. Zafeiriou, M. Pantic, and D. Rueckert. A dynamic approach to the recognition of 3d facial expressions and their temporal models. In *Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition (FG'11), Special Session: 3D Facial Behavior Analysis and Understanding*, Santa Barbara, CA, USA, March 2011.

[4] G. Sandbach, S. Zafeiriou, M. Pantic, and D. Rueckert. Recognition of 3d facial expression dynamics. *Image and Vision Computing*, February 2012. (in press).

[5] A. Srivastava, E. Klassen, S. Joshi, and I. Jermyn. Shape analysis of elastic curves in euclidean spaces. *To appear, IEEE Transactions on, Pattern Analysis and Machine Intelligence*, 2011.

[6] Y. Sun, X. Chen, M. Rosato, and L. Yin. Tracking vertex flow and model adaptation for three-dimensional spatiotemporal face analysis. *Trans. Sys. Man Cyber. Part A*, 40:461–474, May 2010.

[7] Y. Sun and L. Yin. Facial expression recognition based on 3d dynamic range model sequences. In *Proceedings of the 10th European Conference on Computer Vision: Part II*, ECCV '08, pages 58–71, 2008.