A Novel Space-Time Representation on the Positive Semidefinite Cone for Facial Expression Recognition

Anis Kacem, Mohamed Daoudi, Boulbaba Ben Amor, Juan Carlos Alvarez-Paiva

A Novel Space-Time Representation on the Positive Semidefinite Cone for Facial Expression Recognition

Anis Kacem[1], Mohamed Daoudi[1], Boulbaba Ben Amor[1], and Juan Carlos Alvarez-Paiva[2]

[1]IMT Lille Douai, Univ. Lille, CNRS, UMR 9189 – CRIStAL –
Centre de Recherche en Informatique Signal et Automatique de Lille, F-59000 Lille, France
[2]Univ. Lille, CNRS, UMR 8524, Laboratoire Paul Painlevé, F-59000 Lille, France.

Abstract

In this paper, we study the problem of facial expression recognition using a novel space-time geometric representation. We describe the temporal evolution of facial landmarks as parametrized trajectories on the Riemannian manifold of positive semidefinite matrices of fixed-rank. Our representation has the advantage to bring naturally a second desirable quantity when comparing shapes – the spatial covariance – in addition to the conventional affine-shape representation. We derive then geometric and computational tools for rate-invariant analysis and adaptive re-sampling of trajectories, grounding on the Riemannian geometry of the manifold. Specifically, our approach involves three steps: 1) facial landmarks are first mapped into the Riemannian manifold of positive semidefinite matrices of rank 2, to build time-parameterized trajectories; 2) a temporal alignment is performed on the trajectories, providing a geometry-aware (dis-)similarity measure between them; 3) finally, pairwise proximity function SVM (ppfSVM) is used to classify them, incorporating the latter (dis-)similarity measure into the kernel function. We show the effectiveness of the proposed approach on four publicly available benchmarks (CK+, MMI, Oulu-CASIA, and AFEW). The results of the proposed approach are comparable to or better than the state-of-the-art methods when involving only facial landmarks.

1. Introduction

In recent years, Automated Facial Expression Recognition (AFER) has aroused considerable interest [8]. Earlier literature mostly focused on static faces grounding on either shape (geometry) or appearance features. Recently, there have been a general shift to exploit the dynamics (motion) in facial videos [14, 20, 25], as conveying an expression is obviously a temporal process. In particular, advances in landmarks detection [3, 34, 44] have opened the door to accurate geometry-driven approaches. Besides, it has been stated that in unconstrained scenario, geometric features outperform appearance features [23]. However, analyzing temporal shape features brings new challenges – (1) which suitable representation to facial shape analysis under rigid transformations due to changes in head position and orientation? (2) Which temporal representation for modeling the dynamic of facial expression? (3) How to compare and classify temporal sequences for the purpose of facial expression recognition? To tackle these challenges, we introduce in this work a comprehensive geometric framework which involves the temporal evolution of facial landmarks. Our framework incorporates a novel shape representation using Gramian matrices derived from centered facial landmark configurations and its extension to time-parametrized trajectories on the positive semidefinite cone. We use then appropriate tools to compare and classify trajectories in a rate-invariant fashion, grounding on the geometry of the manifold of interest.

2. Related Work

In the *appearance-based (A)* category, first works extend conventional local features such us SIFT, LBP, and HOG to suit video-based data, giving rise to 3D SIFT [33], LBP-TOP [46], and 3D HOG [21]. In [25], the authors exploit the dynamics of facial expressions and propose a semantics-aware representation. They model a video clip as a Spatio-Temporal Manifold (STM) spanned by local spatio-temporal features called *Expressionlets* built from low-level appearance features. These features are based on clustering cuboids of pre-defined sizes extracted from facial sequences in order to model the manifold of facial expression variations. A temporal alignment among STMs is performed to allow a rate-invariant analysis of facial expressions. Deep Networks based on appearance features have been recently applied on facial image sequences for the pur-

pose of AFER. Elaiwat *et al*. [14] propose a restricted Boltzmann machine (RBM) network that, unlike typical deep models, is shallow and therefore easier to optimize. The key property of the RBM network is to disentangle expression-related image transformations from transformations that are not related to the expressions. Despite their investigation in expression recognition, deep networks are less effective if trained with small datasets [37]. To overcome this limitation, Jung *et al*. exploit two temporal features from the appearance and the geometry (landmarks) to train two deep networks termed respectively DTAN and DTGN [20]. They are then combined using a joint fine-tuning method to give rise to the DTAGN. Finally, it has been shown in [36] that face analysis using deep networks is sensitive to pose variations and often requires a face alignment step. As far as the *geometry-based (G)* approaches are concerned, in [18], the authors propose a probabilistic method to capture the subtle motions within expressions using Latent-Dynamic Conditional Random Fields (LDCRFs) on both geometric and appearance features. They illustrate experimentally that variations in shape are much more important than appearance for AFER. In another work, Wang *et al*. [43] introduce a unified probabilistic framework based on an interval temporal Bayesian network (ITBN) built from the movements of specific geometric points detected on the face along a sequence. Recently, shape trajectory-based methods showed their effectiveness in many temporal pattern recognition tasks, especially in action recognition [2, 4, 6, 9, 40]. Taheri *et al*. [35] propose an affine-invariant shape representation on the Grassmann manifold $\mathcal{G}(2, n)$ [5] and model the dynamic of facial expression by parametrized trajectories on this manifold. Geodesic velocities between facial shapes are then used to capture the facial deformations. The classification was achieved using LDA followed by SVM.

From the discussion above, we propose a novel shape representation invariant to rigid motions by embedding shapes into a Positive Semidefinite Riemannian manifold. Facial expression sequences are then viewed as trajectories on this manifold. To compare and classify these trajectories, we propose a variant of SVM that takes into account the nonlinearity of this space. The full approach is illustrated in Fig.1. In summary, the main contributions of this paper are:

- A novel static shape representation based on computing the Gramian matrix from centered landmark configurations, as well as a comprehensive study of the Riemannian geometry of the space of representations (called the cone of Positive Semidefinite $n \times 2$ matrices of fixed-rank 2). Despite the large use of these matrices in several research fields, to the best of our knowledge, this is the first application in static and dynamic shape analysis.
- A temporal extension of the representation via parametrized trajectories in the underlying Riemannian manifold, with associated computational tools for temporal alignment and adaptive re-sampling of trajectories.
- The classification of trajectories based on pairwise proximity function SVM (ppfSVM) grounding on pairwise (dis-)similarity measures between them, with respect to the metric of the underlying manifold.
- Extensive experiments and baselines on four publicly datasets and a comparative study with existing literature, which demonstrates the competitiveness of the approach.

The rest of the paper is organized as following. In section 3, we study the Riemannian geometry of the Positive Semidefinite manifold. In section 4, we adopt a temporal extension of the representation via time-parametrized trajectories in the manifold, with the definition of relevant geometric tools for temporal registration and trajectory re-sampling. Section 5 states the classification approach based on a variant of the standard SVM associated to a closeness between the trajectories. Experimental results and discussions are reported in section 6. In section 7, we conclude and draw some perspectives of the work.

3. Shape and Trajectory Representations

Let us consider $\{Z_0, \ldots, Z_N\}$ an arbitrary sequence of landmark configurations. Each configuration Z_i $(0 \leq i \leq N)$ is an $n \times 2$ matrix $[(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)]^T$ of rank 2 encoding the positions of n distinct points on the plane: $p_1 = (x_1, y_1), \ldots, p_n = (x_n, y_n)$. We are interested in studying such sequences or curves of landmark configurations up to Euclidean motions of the plane. In what follows, we will first study static observations representation, then adopt a time-parametrized representation for a temporal analysis.

As a first step, we seek a shape representation that is invariant up to Euclidean transformations (rotation and translation). Arguably, the most natural choice is the matrix of pairwise distances between the landmark points of the same shape augmented by the distances from all landmarks to the center of mass $p_0 = (\bar{x}, \bar{y})$. Since we are dealing with Euclidean distances, it will turn out to be more convenient to consider the matrix of the squares of these distances. Also note that by subtracting the center of mass from the coordinates of the landmarks, these can be considered as *centered*: the center of mass is always at the origin. From now on we will assume $p_0 = (\bar{x}, \bar{y}) = (0, 0)$. With this provision, the augmented pairwise square-distance matrix $\mathcal{D}$ takes the form,

$$\mathcal{D} := \begin{pmatrix} 0 & \|p_1\|^2 & \cdots & \|p_n\|^2 \\ \|p_1\|^2 & 0 & \cdots & \|p_1 - p_n\|^2 \\ \vdots & \vdots & \vdots & \vdots \\ \|p_n\|^2 & \|p_n - p_1\|^2 & \cdots & 0 \end{pmatrix},$$

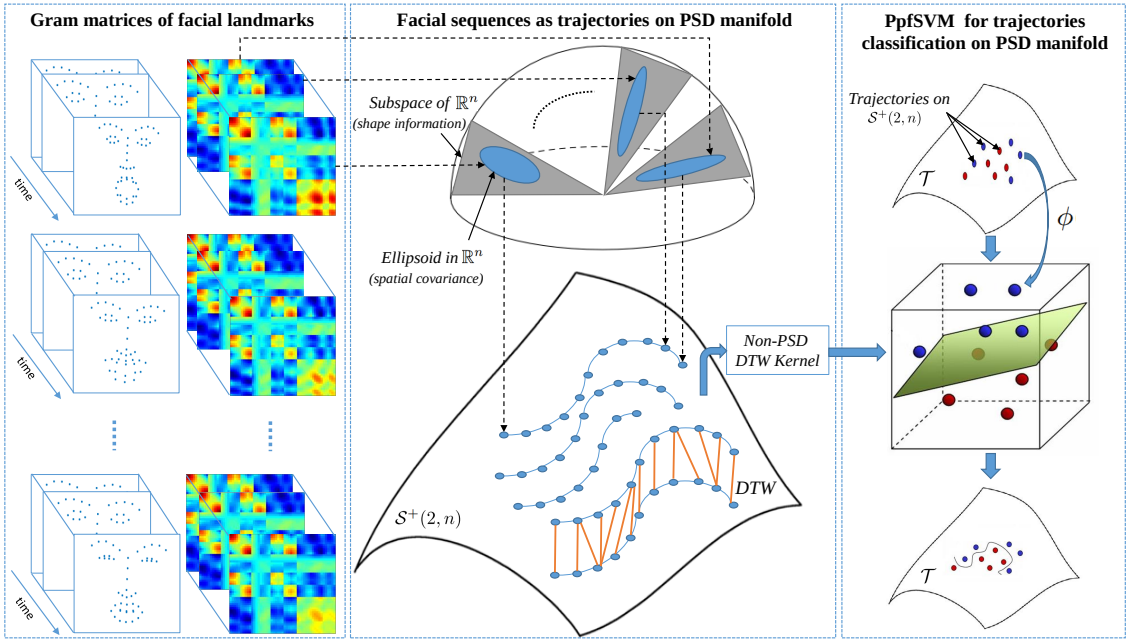

Figure 1. Overview of the proposed approach – After automatic landmark detection for each frame of the video, the Gram matrices are computed to build trajectories on $\mathcal{S}^+(2,n)$. A moving shape is hence assimilated to an ellipsoid traveling along 2-dimensional subspaces of $\mathbb{R}^n$ with $d_{\mathcal{S}^+}$ to compare static ellipsoids. Dynamic Time Warping (DTW) is then used to align and compare trajectories in a rate-invariant manner. Finally, the ppfSVM is used on these trajectories as expression classifier.

where $p_i = (x_i, y_i)$ for all $1 \leq i \leq n$. As usual, $\|\cdot\|$ denotes the norm associated to the l^2-inner product $\langle \cdot, \cdot \rangle$.

A key observation is that the matrix $\mathcal{D}$ can be easily read from the $n \times n$ Gram matrix $G := ZZ^T$. Indeed, the entries of G are the pairwise inner products of the points $p_1, \ldots, p_n$,

$$G = ZZ^T = \langle p_i, p_j \rangle \;\; 1 \leq i, j \leq n \,, \tag{1}$$

and the equality

$$\mathcal{D}_{ij} = \langle p_i, p_i \rangle - 2\langle p_i, p_j \rangle + \langle p_j, p_j \rangle \;\; (0 \leq i, j \leq n), \tag{2}$$

establishes a linear equivalence between the set of $n \times n$ Gram matrices and augmented square-distance $(n+1) \times (n+1)$ matrices of distinct points on the plane. On the other hand, Gram matrices of the form ZZ^T, where Z is an $n \times 2$ matrix of rank 2, are characterized as $n \times n$ positive semidefinite matrices of rank 2 (for a detailed account of the relation between positive semidefinte matrices, Gram matrices, and square-distance matrices, we refer the reader to Section 6.2.1 of the book [10]). Conveniently for us, the Riemannian geometry of the space of these matrices, called the positive semidefinite cone $\mathcal{S}^+(2,n)$, was studied in [7, 15, 27, 38].

An alternative shape representation considered in [5] and [35] associates to each configuration Z the two-dimensional subspace $span(Z)$ spanned by its columns. This representation, which exploits the well-known geometry of the Grassmann manifold $\mathcal{G}(2,n)$ of two-dimensional subspaces in $\mathbb{R}^n$, is invariant under *all* invertible linear transformations. By fully encoding the set of all mutual distances between landmark points, the Euclidean shape representation proposed in this paper supplements the affine shape representation with the knowledge of the 2×2 covariance matrix for the centered landmarks. This leads to considerable improvements in the results of the conducted facial expression recognition experiments.

3.1. Riemannian geometry of $\mathcal{S}^+(2,n)$

Given an $n \times 2$ matrix Z of rank two, its polar decomposition $Z = UR$ with $R = (Z^TZ)^{1/2}$ allows us to write the Gram matrix ZZ^T as UR^2U^T. Since the columns of the matrix U are orthonormal, this decomposition defines a map

$$\begin{aligned} \Pi : & V_{n,2} \times \mathcal{P}_2 \to \mathcal{S}^+(2,n) \\ & (U, R^2) \mapsto UR^2U^T \end{aligned}$$

from the product of the Stiefel manifold $V_{n,2}$ and the cone of 2×2 positive definite matrices $\mathcal{P}_2$ to the manifold $\mathcal{S}^+(2,n)$ of $n \times n$ positive semidefinite matrices of rank two. The map Π defines a principal fiber bundle over $\mathcal{S}^+(2,n)$ with fibers

$$\Pi^{-1}(UR^2U^T) = \{(UO, O^TR^2O) : O \in \mathcal{O}(2)\} \,,$$

where $\mathcal{O}(2)$ is the group of 2×2 orthogonal matrices. Bonnabel and Sepulchre [7] use this map and the geometry

of the *structure space* $V_{n,2} \times \mathcal{P}_2$ to introduce a Riemannian metric on $\mathcal{S}^+(2,n)$ and study its geometry.

3.2. Tangent space and Riemannian metric

The tangent space $T_{(U,R^2)}(V_{n,2} \times \mathcal{P}_2)$ consists of pairs (M,N), where M is a $n \times 2$ matrix satisfying $M^TU + U^TM = 0$ and N is any 2×2 symmetric matrix. Bonnabel and Sepulchre define a *connection* (see [22, p. 63]) on the principal bundle $\Pi : V_{n,2} \times \mathcal{P}_2 \to \mathcal{S}^+(2,n)$ by setting the horizontal subspace $\mathcal{H}_{(U,R^2)}$ at the point (U,R^2) to be the space of tangent vectors (M,N) such that $M^TU = 0$ and N is an arbitrary 2×2 symmetric matrix. They also define an inner product on $\mathcal{H}_{(U,R^2)}$: given two tangent vectors $A = (M_1,N_1)$ and $B = (M_2,N_2)$ on $\mathcal{H}_{(U,R^2)}$, set

$$\langle (A,B) \rangle_{\mathcal{H}_{U,R^2}} = \mathrm{tr}(M_1^TM_2) + k\,\mathrm{tr}(N_1R^{-2}N_2R^{-2}), \quad (3)$$

where $k > 0$ is a real parameter.

It is easily checked that the action of the group of 2×2 orthogonal matrices on the fiber $\Pi^{-1}(UR^2U^T)$ sends horizontals to horizontals isometrically. It follows that the inner product on $T_{UR^2U^T}\mathcal{S}^+(2,n)$ induced from that of $\mathcal{H}_{(U,R^2)}$ via the linear isomorphism $D\Pi$ is independent of the choice of point (U,R^2) projecting onto UR^2U^T. This procedure defines a Riemannian metric on $\mathcal{S}^+(2,n)$ for which the natural projection

$$\rho : \mathcal{S}^+(2,n) \to \mathcal{G}(2,n)$$
$$G \mapsto range(G)$$

is a Riemannian submersion. This allows us to relate the geometry of $\mathcal{S}^+(2,n)$ with that of the Grassmannian $\mathcal{G}(2,n)$.

Recall that the geometry of the Grassmannian $\mathcal{G}(2,n)$ is easily described by using the map

$$span : V_{n,2} \to \mathcal{G}(2,n)$$

that sends an $n \times 2$ matrix with orthonormal columns U to their span $span(U)$. Given two subspaces $\mathcal{U}_1 = span(U_1)$ and $\mathcal{U}_2 = span(U_2) \in \mathcal{G}(2,n)$, the geodesic curve connecting them is

$$span(U(t)) = span(U_1\cos(\Theta t) + M\sin(\Theta t)), \quad (4)$$

where $\Theta = diag(\theta_1,\theta_2)$ is a 2×2 diagonal matrix formed by the *principal angles* between $\mathcal{U}_1$ and $\mathcal{U}_2$, while the matrix M is given by the formula $M = (I_n - U_1U_1^T)U_2F$, with F being the pseudoinverse $diag(\sin(\theta_1),\sin(\theta_2))$. The Riemannian distance between $\mathcal{U}_1$ and $\mathcal{U}_2$ is given by

$$d^2_{\mathcal{G}}(\mathcal{U}_1,\mathcal{U}_2) = \|\Theta\|^2_F. \quad (5)$$

3.3. Pseudo-geodesics and closeness in $\mathcal{S}^+(2,n)$

Bonnabel and Sepulchre ([7]) define the *pseudo-geodesic* connecting two matrices $G_1 = U_1R_1^2U_1^T$ and $G_2 = U_2R_2^2U_2^T$ in $\mathcal{S}^+(2,n)$ as the curve

$$\mathcal{C}_{G_1 \to G_2}(t) = U(t)R^2(t)U^T(t), \forall t \in [0,1], \quad (6)$$

where $R^2(t) = R_1\exp(t\log R_1^{-1}R_2^2R_1^{-1})R_1$ is a geodesic in $\mathcal{P}_2$ and $U(t)$ is the geodesic in $\mathcal{G}(2,n)$ given by Eq.(4). They also define the *closeness* between G_1 and G_2, $d_{\mathcal{S}^+}(G_1,G_2)$, as the square of the length of this curve:

$$\begin{aligned} & d_{\mathcal{S}^+}(G_1,G_2) \\ & = \|\Theta\|^2_F + k\|\log R_1^{-1}R_2^2R_1^{-1}\|^2_F \\ & = d^2_{\mathcal{G}}(span(U_1),span(U_2)) + kd^2_{\mathcal{P}_2}(R_1^2,R_2^2)\,. \end{aligned} \quad (7)$$

The closeness $d_{\mathcal{S}^+}$ consists of two independent contributions, the square of the distance $d_{\mathcal{G}}(span(U_1),span(U_2))$ between the two associated subspaces and the square of the distance $d_{\mathcal{P}_2}(R_1^2,R_2^2)$ on the positive cone $\mathcal{P}_2$ (Fig.2). Note that $\mathcal{C}_{G_1 \to G_2}$ is not necessarily a geodesic and therefore, the closeness $d_{\mathcal{S}^+}$ is not a true Riemannian distance. From the viewpoint of the landmark configurations Z_1 and Z_2, with $G_1 = Z_1Z_1^T$ and $G_2 = Z_2Z_2^T$, the closeness encodes the distances measured between the affine shapes $span(Z_1)$ and $span(Z_2)$ in $\mathcal{G}(2,n)$ and between their spatial covariances in $\mathcal{P}_2$. Indeed, the spatial covariance of Z_i $(i = 1,2)$ is the 2×2 symmetric positive definite matrix

$$C = \frac{Z_i^TZ_i}{n-1} = \frac{(U_iR_i)^T(U_iR_i)}{n-1} = \frac{R_i^2}{n-1}. \quad (8)$$

The weight parameter k controls the relative contribution of these two informations. Note that for $k = 0$ the distance on $\mathcal{S}^+(2,n)$ collapses to the distance on $\mathcal{G}(2,n)$. Nevertheless, the authors in [7] recommend choosing small values for this parameter. The conducted experiments for expression recognition reported in section 6 are in accordance with this recommendation.

For more details about the geometry of the Grassmannians $\mathcal{G}(2,n)$ and the positive define cone $\mathcal{P}_2$, readers are referred to [1, 5, 7, 28].

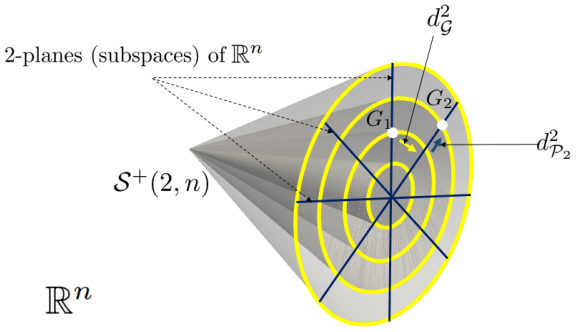

Figure 2. A pictorial representation of the positive semidefinite cone $\mathcal{S}^+(2,n)$. Viewing matrices G_1 and G_2 as ellipsoids in $\mathbb{R}^n$, the closeness consists of two contributions: $d^2_{\mathcal{G}}$ (squared Grassmann distance) and $d^2_{\mathcal{P}_2}$ (squared Riemannian distance in $\mathcal{P}_2$).

4. Modeling Facial Expressions as Trajectories in $\mathcal{S}^+(2,n)$

We are able to compare static shape representations based on their Gramian representation G, the induced space, and closeness introduced in the previous section. We need a natural and effective extension to study their temporal evolution. Following [6, 35, 39], we define curves $\beta_G : I \rightarrow \mathcal{S}^+(2,n)$ (I denotes the time domain, *e.g.* $[0,1]$) to model the spatio-temporal evolution of elements on $\mathcal{S}^+(2,n)$. Given a sequence of shapes $\{Z_0,\ldots,Z_N\}$ represented by their corresponding Gram matrices $\{G_0,\ldots,G_N\}$ in $\mathcal{S}^+(2,n)$, the corresponding curve is the trajectory of the point $\beta_G(t)$ on $\mathcal{S}^+(2,n)$, when t ranges in $[0,1]$. These curves are obtained by connecting all successive Gramian representations of shapes G_i and G_{i+1}, $0 \leq i \leq N-1$, by pseudo-geodesics in $\mathcal{S}^+(2,n)$.

4.1. Temporal alignment and analysis

The execution rate of facial expressions is often arbitrary and that results in different parameterizations of corresponding trajectories. This parameterization variability distorts the comparison measures of these trajectories. Given m trajectories $\{\beta_G^1,\beta_G^2,\ldots,\beta_G^m\}$ on $\mathcal{S}^+(2,n)$, we are interested in finding functions γ_i such that the $\beta_G^i(\gamma_i(t))$ are matched optimally for all $t \in [0,1]$. In other words, two curves $\beta_G^1(t)$ and $\beta_G^2(t)$ represent the same trajectories if their images are the same. This happens if, and only if, $\beta_G^2 = \beta_G^1 \circ \gamma$, where γ is a re-parameterization of the interval $[0,1]$. The problem of temporal alignment is turned to find an optimal warping function $\gamma^\star$ according to,

$$\gamma^\star = \arg\min_{\gamma\in\Gamma} \int_0^1 d_{\mathcal{S}^+}(\beta_G^1(t),\beta_G^2(\gamma(t)))\,\mathrm{d}t\,, \qquad (9)$$

where Γ denotes the set of all monotonically-increasing functions $\gamma : [0,1] \rightarrow [0,1]$. The most commonly used method to solve such optimization problem is the Dynamic Time Warping (DTW) algorithm. Note that accommodation of the DTW algorithm to the manifold-values sequences can be achieved with respect to an appropriate metric defined on the underlying manifold $\mathcal{S}^+(2,n)$. Having the optimal re-parametrization function $\gamma^\star$, one can define a (dis-)similarity measure between two trajectories allowing a rate-invariant comparison:

$$d_{DTW}(\beta_G^1,\beta_G^2) = \int_0^1 d_{\mathcal{S}^+}(\beta_G^1(t),\beta_G^2(\gamma^\star(t)))\,\mathrm{d}t. \quad (10)$$

From now, we shall use $d_{DTW}(.,.)$ to compare trajectories in our manifold of interest $\mathcal{S}^+(2,n)$.

4.2. Adaptive re-sampling of trajectories

One difficulty in video analysis is to catch the most relevant frames and focus on them. In fact, it is relevant to reduce the number of frames when no motion happened and in the same time "introduce" new frames, otherwise. Our geometric framework provides tools to do so. In fact, interpolation between successive frames could be achieved using pseudo-geodesics defined in Eq.(6), while their length (closeness defined in Eq.(7)) expresses the magnitude of the motion. Accordingly, we have designed an adaptive resampling tool that is able to increase/decrease the number of samples in a fixed time interval according to their relevance, with respect to the geometry of the underlying manifold $\mathcal{S}^+(2,n)$. Relevant samples are identified by a relatively low closeness $d_{\mathcal{S}^+}$ to the previous frame, while irrelevant ones correspond to a higher closeness level. Here, the down-sampling is performed by removing irrelevant shapes. In turn, the up-sampling is possible by interpolating between successive shape representations in $\mathcal{S}^+(2,n)$ using pseudo-geodesics.

More formally, given a trajectory $\beta_G(t)_{t=0,1,\ldots,N}$ on $\mathcal{S}^+(2,n)$, for each sample $\beta_G(t)$ we compute the closeness to the previous sample, *i.e.* $d_{\mathcal{S}^+}(\beta_G(t),\beta_G(t-1))$. If the value is below a defined threshold ζ_1, current sample is simply removed from the trajectory. In contrast, if the distance exceeds a second threshold ζ_2, new samples (shapes) generated from the pseudo-geodesic curve connecting $\beta_G(t)$ to $\beta_G(t-1)$ are inserted in the trajectory.

5. Classification of Trajectories in $\mathcal{S}^+(2,n)$

Our trajectory representation reduces the problem of facial sequences classification to trajectories classification in $\mathcal{S}^+(2,n)$. Let us consider $\mathcal{T} = \{\beta_G : [0,1] \rightarrow \mathcal{S}^+(2,n)\}$, the set of time-parametrized trajectories of the underlying manifold. Let $\mathcal{L} = \{(\beta_G^1,y^1),\ldots,(\beta_G^m,y^m)\}$ be the training set with class labels, where $\beta_G^i \in \mathcal{T}$ and $y^i \in \mathcal{Y}$, *e.g.* $\mathcal{Y} = \{(\text{Ha}),(\text{An})\}$, such that $y^i = f(\beta_G^i)$. The goal here is to find an approximation h to f such that $h : \mathcal{T} \rightarrow \mathcal{L}$. In Euclidean spaces, any standard classifier (*e.g.* standard SVM) may be a natural and appropriate choice to classify the trajectories. Unfortunately, this is no more suitable as the space $\mathcal{T}$ built from $\mathcal{S}^+(2,n)$ is non-linear. A function that divides the manifold is rather a complicated notion compared with the Euclidean space. In current literature, few approaches have been proposed to handle the nonlinearity of Riemannian manifolds [19, 35, 39, 40]. These methods map the points on the manifold to a tangent space or to Hilbert space where traditional learning techniques can be used for classification. Mapping data to a tangent space only yields a first-order approximation of the data that can be distorted, especially in regions far from the origin of the tangent space. Moreover, iteratively mapping back and forth, *i.e.* Riemannian Logarithmic and exponential maps, to the tangent spaces significantly increases the computational cost of the algorithm. Recently, some authors propose to embed a manifold in a high dimensional Reproducing Kernel Hilbert Space (RKHS), where Euclidean geome-

try can be applied [19]. The Riemannian kernels enable the classifiers to operate in an extrinsic feature space without computing tangent space and $\log$ and $\exp$ maps. Many Euclidean machine learning algorithms can be directly generalized to an RKHS, which is a vector space that possesses an important structure: the inner product. Such an embedding, however, requires a positive semi-definite kernel function, according to Mercer's theorem [32].

Inspired by a recent work of [4] for action recognition, we adopt the *pairwise proximity function SVM* (ppfSVM) [16, 17]. PpfSVM requires the definition of a (dis-)simlarity measure to compare samples. In our case, it is natural to consider the d_{DTW} defined in Eq.(10) for such a comparison. This strategy involves the construction of inputs such that each trajectory is represented by its (dis-)similarity to all the trajectories in the dataset, with respect to d_{DTW}, and then apply a conventional SVM to this transformed data [17]. The ppfSVM is related to the arbitrary kernel-SVM without restrictions on the kernel function [16].

Given m trajectories $\{\beta_G^1, \beta_G^2, \ldots, \beta_G^m\}$ in $\mathcal{T}$, we follow [4] and define a proximity function $\mathcal{P}_{\mathcal{T}} : \mathcal{T} \times \mathcal{T} \rightarrow \mathbb{R}_+$ between two trajectories $\beta_G^1, \beta_G^2 \in \mathcal{T}$ as following,

$$\mathcal{P}_{\mathcal{T}}(\beta_G^1, \beta_G^2) = d_{DTW}(\beta_G^1, \beta_G^2). \tag{11}$$

According to [16], there are no restrictions on the function $\mathcal{P}_{\mathcal{T}}$. For an input trajectory $\beta_G \in \mathcal{T}$, the mapping $\phi(\beta_G)$ is given by,

$$\phi(\beta_G) = [\mathcal{P}_{\mathcal{T}}(\beta_G, \beta_G^1), \ldots, \mathcal{P}_{\mathcal{T}}(\beta_G, \beta_G^m)]^T. \tag{12}$$

The obtained vector $\phi(\beta_G) \in \mathbb{R}^m$ is used to represent a sample trajectory $\beta_G \in \mathcal{T}$. Hence, the set of trajectories can be represented by a $m \times m$ matrix P, where $P(i,j) = \mathcal{P}_{\mathcal{T}}(\beta_G^i, \beta_G^j)$, $i, j \in \{1, \ldots, m\}$. Finally, a linear SVM is applied to this data representation. Further details on ppfSVM can be found in [4, 16, 17].

6. Experimental Results

To validate the proposed approach, we have conducted extensive experiments on four publicly available datasets – CK+, MMI, Oulu-CASIA, and AFEW. We have followed experimental settings commonly used in recent works. Note that all our experiments are made once the facial landmarks are extracted using the method proposed in [3] on CK+, MMI, and Oulu-CASIA datasets. On the challenging AFEW, we have considered the corrections provided in [1] after applying the same detector.

Cohn-Kanade Extended (CK+) database [26] – is one of the most popular datasets. It contains 123 subjects and 593 frontal image sequences of posed expressions. Among them, 118 subjects are annotated with the seven labels – anger (An), contempt (Co), disgust (Di), fear (Fe), happy (Ha), sad (Sa), and surprise (Su). Note that only the two first temporal phases of the expression, *i.e*. neutral and onset (with apex frames), are present. Following the same settings of [14, 20], we have performed 10-fold cross validation experiment. The results are summarized in Table 1.

Table 1. Confusion matrix of the proposed trajectory representation and classification on $\mathcal{S}^+(2,n)$ – CK+ database.

	An	Co	Di	Fe	Ha	Sa	Su
An	**100**	5.55	3.38	0	0	3.57	0
Co	0	**83.35**	0	0	1.44	0	1.2
Di	0	0	**96.62**	0	0	0	0
Fe	0	0	0	**92**	0	0	0
Ha	0	5.55	0	8	**98.56**	0	0
Sa	0	5.55	0	0	0	**96.43**	0
Su	0	0	0	0	0	0	**98.8**

Overall, the average accuracy is 96.87%. When individual accuracy of (An), (Di), (Ha), and (Su) are high, recognizing (Co) and (Fe) is still challenging. Note that the accuracy of the trajectory representation on $\mathcal{G}(2,n)$, following the same pipeline is 2% lower, which confirms the contribution of the covariance embedded to our shape representation.

MMI database [37] – consists of 205 image sequences with frontal faces of only 30 subjects labeled with the six basic emotion labels. This database is different from the other databases; each sequence begins with a neutral facial expression and has a posed facial expression in the middle of the sequence. This ends with the neutral facial expression. The location of the peak frame is not provided as a prior information. Here again, the protocol used in [14, 20] was followed according to a 10-fold cross-validation schema. The confusion matrix is reported in Table 2. An average classification accuracy of 79.19% is reported. Note that based on geometric features only, our approach grounding on both representations on $\mathcal{S}^+(2,n)$ and $\mathcal{G}(2,n)$ achieves competitive results with respect to the literature (see Table 5).

Table 2. Confusion matrix of the proposed trajectory representation and classification on $\mathcal{S}^+(2,n)$ – MMI database.

	An	Di	Fe	Ha	Sa	Su
An	**76.66**	9.37	0	0	9.37	0
Di	13.33	**75**	13.79	2.44	3.12	0
Fe	0	3.12	**55.17**	0	9.37	12.82
Ha	0	12.5	0	**97.56**	0	0
Sa	10	0	3.44	0	**71.87**	2.56
Su	0	0	27.58	0	6.25	**84.61**

Oulu-CASIA database [45] – includes 480 image sequences of 80 subjects taken under normal illumination conditions. They are labeled with one of the six basic emotion labels. Each sequence begins with a neutral facial expression and ends with the apex of the expression. We adopt a 10-fold cross validation schema similarly to [20, 25]. This

[1] http://sites.google.com/site/chehrahome

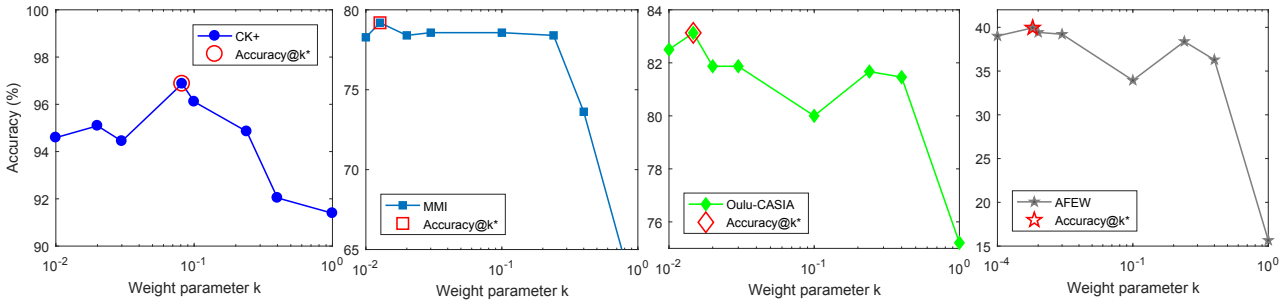

Figure 3. Accuracy of the proposed approach when varying the weight parameter k, on (left to right) CK+, MMI, Oulu-CASIA, and AFEW.

time, the average accuracy is 83.13%, hence 3% higher than the Grassmann trajectory representation. This is the highest accuracy reported in literature (refer to Table 6).

Table 3. Confusion matrix of the proposed trajectory representation and classification on $\mathcal{S}^+(2,n)$ – Oulu-CASIA database.

	An	Di	Fe	Ha	Sa	Su
An	**81.25**	15	1.25	0	13.75	0
Di	10	**78.75**	2.5	0	6.25	0
Fe	1.25	1.25	**78.75**	6.25	3.75	5
Ha	1.25	1.25	3.75	**91.25**	1.25	1.25
Sa	6.25	3.75	5	2.5	**75**	0
Su	0	0	8.75	0	0	**93.75**

AFEW database [12] – collected from movies showing close-to-real-world conditions, which depicts or simulates the spontaneous expressions in uncontrolled environment. According to the protocol defined in EmotiW'2013 [11], the database is divided into three sets: training, validation, and test. The task is to classify each video clip into one of the seven expression categories (the six basic emotions plus the neutral). As the ground truth of test set is still unreleased, here we only report our results on the validation set for comparison with [11, 14, 25]. The average accuracy is 39.94%. Unsurprisingly, the (Ne), (An), and (Ha) are better recognized over the rest. Despite their competitiveness with respect to recent literature, these results state clearly that the AFER "in-the-wild" is still a distant goal.

Table 4. Confusion matrix of the proposed trajectory representation and classification on $\mathcal{S}^+(2,n)$ – AFEW database following the EmotiW'13 protocol [11].

	An	Di	Fe	Ha	Ne	Sa	Su
An	**56.25**	12.5	30.43	4.76	7.93	13.11	26.08
Di	0	**10**	8.69	4.76	0	6.55	2.17
Fe	7.81	7.5	**26.08**	4.76	7.93	14.75	19.56
Ha	10.93	22.5	10.87	**66.66**	6.35	11.47	2.17
Ne	9.37	37.5	10.87	12.69	**63.49**	32.78	30.43
Sa	10.93	2.5	6.52	6.35	11.11	**18.03**	2.17
Su	4.68	7.5	6.52	0	3.17	3.27	**17.39**

In Fig.3 we study the method's behavior when varying the parameter k (of the closeness) defined in Eq.(7). Recall that k serves to balance the contribution of the distance between covariance matrices living in $\mathcal{P}_2$ with respect to the Grassmann contribution $\mathcal{G}(2,n)$. The graphs report the method accuracy respectively on CK+, MMI, Oulu-CASIA, and AFEW. The optimal performances are achieved for the following values – $k^*_{CK+} = 0.081$, $k^*_{MMI} = 0.012$, $k^*_{Oulu-CASIA} = 0.014$, and $k^*_{AFEW} = 0.001$.

Table 5. Overall accuracy (%) on CK+ and MMI datasets. Here, $^{(A)}$: appearance; $^{(G)}$: geometry (or shape); ^{s}: static; *: Deep Learning based approach; last raw: ours

Method	**CK+**	**MMI**
$^{(A)}$ 3D HOG [21] (from [20])	91.44	60.89
$^{(A)}$ 3D SIFT [33] (from [20])	-	64.39
$^{(A)}$ Cov3D [30] (from [20])	92.3	-
$^{(A)}$ MSR [29] (LOSO)	91.4	-
$^{(A)}$ STM-ExpLet [25] (10-fold)	**94.19**	**75.12**
$^{(A)}$ CSPL [48] (10-fold)	89.89	73.53
$^{(A)}$ F-Bases [31] (LOSO)	**96.02**	**75.12**
$^{(A)}$ ST-RBM [14] (10-fold)	**95.66**	**81.63**
$^{(A)}$ FaceNet2ExpNet^{*,s} [13]	**96.8**	-
$^{(A)}$ 3DCNN-DAP [24] * (15-fold)	87.9	62.2
$^{(A)}$ DTAN [20] * (10-fold)	91.44	62.45
$^{(A+G)}$ DTAGN [20] * (10-fold)	**97.25**	**70.24**
$^{(G)}$ DTGN [20] * (10-fold)	92.35	59.02
$^{(G)}$ TMS [18] (4-fold)	85.84	-
$^{(G)}$ HMM [43] (15-fold)	83.5	51.5
$^{(G)}$ ITBN [43] (15-fold)	86.3	59.7
$^{(G)}$ Velocity on $\mathcal{G}(n,2)$[35]	82.8	-
$^{(G)}$ traj. on $\mathcal{G}(2,n)$ (10-fold)	94.25 ± 3.71	78.18 ± 4.87
$^{(G)}$ **traj. on $\mathcal{S}^+(2,n)$ (10-fold)**	$\mathbf{96.87 \pm 2.46}$	$\mathbf{79.19 \pm 4.62}$

Comparative study with the state-of-the-art. In tables 5 and 6, we compare our approach over the recent literature. Overall, our approach achieves competitive performances with respect to the most recent approaches. On CK+, we obtained the second highest accuracy. The ranked-first approach is DTAGN [20], in which two deep networks are trained on shape and appearance channels, then fused. Note that the geometry deep network (DTGN) achieved 92.35%, which is much lower than ours. Furthermore, our approach outperforms the ST-RBM [14] and the STM-ExpLet [25]. On MMI dataset, our approach outperforms the DTAGN [20] and the STM-ExpLet [25]. However, it is behind ST-RBM [14]. Note that the FaceNet2ExpNet [13]

is a pure static approach and is reported here as the state-of-the-art of static AFER.

Table 6. Overall accuracy on Oulu-CASIA and AFEW dataset (following the EmotiW'13 protocol [11])

Method	Oulu-CASIA	AFEW
$^{(A)}$ HOG 3D [21]	70.63	26.90
$^{(A)}$ HOE [41]	-	19.54
$^{(A)}$ 3D SIFT [33]	55.83	24.87
$^{(A)}$ LBP-TOP [47]	68.13	25.13
$^{(A)}$ EmotiW [11]	-	27.27
$^{(A)}$ STM [25]	-	29.19
$^{(A)}$ STM-ExpLet [25]	74.59	31.73
$^{(A+G)}$ DTAGN [20] * (10-fold)	**81.46**	-
$^{(A)}$ ST-RBM [14]	-	**46.36**
$^{(G)}$ **traj. on** $\mathcal{G}(2,n)$	80.0 ± 5.22	39.1
$^{(G)}$ **traj. on** $\mathcal{S}^+(2,n)$	$\mathbf{83.13 \pm 3.86}$	**39.94**

On Oulu-CASIA dataset, our approach shows a clear superiority to existing methods, in particular STM-ExpLet [25] and DTGN [20]. Elaiwat *et al*. [14] do not report any results on this dataset. However, their approach achieved the highest accuracy on AFEW. Our approach is ranked second showing a superiority to remaining approaches on AFEW.

Baseline experiments. Based on the results reported in table 7, we discuss in this paragraph algorithms and their computational complexity with respect to baselines.

Table 7. Baseline experiments on CK+, MMI, and AFEW datasets.

Distance	CK+ (%)	Time (s)
Flat distance $d_{\mathcal{F}^+}$	93.78 ± 2.92	0.020
Distance $d_{\mathcal{P}_n}$ in $\mathcal{P}_n$	92.92 ± 2.45	0.816
Closeness $d_{\mathcal{S}^+}$	$\mathbf{96.87 \pm 2.46}$	0.055

Temporal alignment	CK+ (%)	MMI (%)	Time (s)
without DTW	90.94 ± 4.23	66.93 ± 5.79	0.018
with DTW	$\mathbf{96.87 \pm 2.46}$	$\mathbf{79.19 \pm 4.62}$	0.055

Adaptive re-sampling	MMI (%)	AFEW (%)
without resampling	74.72 ± 5.34	36.81
with resampling	$\mathbf{79.19 \pm 4.62}$	**39.94**

Classifier	CK+ (%)	AFEW (%)
K-NN	88.97 ± 6.14	29.77
ppf-SVM	$\mathbf{96.87 \pm 2.46}$	**39.94**

We highlight firstly the superiority of the trajectory representation on $\mathcal{S}^+(2,n)$ over the Grassmannian (refer to Tables 5 and 6). This is due to the contribution of the covariance part further to the conventional affine-shape analysis over the Grassmannian. Secondly, we have used different distances defined on $\mathcal{S}^+(2,n)$. Specifically, given two matrices G_1 and G_2 in $\mathcal{S}^+(2,n)$: (1) as proposed in [42], we used $d_{\mathcal{P}_n}$ that was defined in Eq.(7) to compare them through regularizing their ranks, *i.e*. making them n full-rank and considering them in $\mathcal{P}_n$ (the space of n-by-n positive definite matrices), $d_{\mathcal{P}_n}(G_1,G_2) = d_{\mathcal{P}_n}(G_1 + \epsilon I_n, G_2 + \epsilon I_n)$; (2) we used the Euclidean flat distance $d_{\mathcal{F}^+}(G_1,G_2) = \|G_1 - G_2\|_F$, where $\|.\|_F$ denotes the Frobenius-norm. The closeness $d_{\mathcal{S}^+}$ between two elements of $\mathcal{S}^+(2,n)$ defined in Eq.(7) is more suitable compared to the distance $d_{\mathcal{P}_n}$ and the flat distance $d_{\mathcal{F}^+}$. This demonstrates the importance of being faithful to the geometry of the manifold of interest. Another advantage of using $d_{\mathcal{S}^+}$ over $d_{\mathcal{P}_n}$ is the computational time as it involves n-by-2 and 2-by-2 matrices instead of n-by-n matrices.

Table 7 reports the average accuracy when DTW is used or not in our pipeline on both CK+ and MMI datasets. It is clear from these experiments that a temporal alignment of the trajectories is a crucial step as an improvement of around 12% is obtained on MMI and 6% on CK+. The adaptive re-sampling tool is also analyzed. When it is involved in the pipeline, an improvement of around 5% is achieved on MMI and 3% on AFEW.

In the last table, we compare the results of ppfSVM to a K-NN classifier for both CK+ and AFEW databases. Each test sequence is classified by a majority vote of its K-nearest neighbors using the (dis-)similarity measure defined in Eq. 10. The number of nearest neighbors K to consider for each database is chosen by cross-validation. On CK+, we obtained an average accuracy of 88.97% for $K = 11$. On AFEW, we obtained an average accuracy of 29.77% for $K = 7$. These results are outperformed by ppfSVM classifier.

7. Conclusion and Future Work

We have proposed in this paper a geometric approach for effectively modeling and classifying dynamic facial sequences. Based on Gramian matrices derived from the facial landmarks, our representation consists of an affine-invariant shape representation and a spatial covariance of the landmarks. We have exploited the geometry of the space to define a closeness between static and dynamic (trajectory) representations. We have derived then computational tools to align, re-sample and compare these trajectories giving rise to a rate-invariant analysis. Finally, facial expressions are learned from these trajectories using a variant of SVM, called ppfSVM, which allows to deal with the nonlinearity of the space of representations. Our experiments on four publicly available datasets showed that the proposed approach gives competitive or better than state-of-art results. In the future, we will extend this approach to handle with smaller variations of facial expressions. Another direction could be adapting our approach for other applications that involve landmark sequences analysis such as action recognition.

8. Acknowledgements

This work has been partially supported by PIA (ANR-11-EQPX-0023), European Founds for the Regional Development (FEDER-Presage 41779).

References

[1] P.-A. Absil, R. Mahony, and R. Sepulchre. Riemannian geometry of grassmann manifolds with a view on algorithmic computation. *Acta Applicandae Mathematica*, 80(2):199–220, 2004.

[2] R. Anirudh, P. K. Turaga, J. Su, and A. Srivastava. Elastic functional coding of riemannian trajectories. *IEEE IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(5):922–936, 2017.

[3] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic. Incremental face alignment in the wild. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pages 1859–1866, 2014.

[4] M. A. Bagheri, Q. Gao, and S. Escalera. Support vector machines with time series distance kernels for action classification. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*, pages 1–7. IEEE, 2016.

[5] E. Begelfor and M. Werman. Affine invariance revisited. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006), 17-22 June 2006, New York, NY, USA*, pages 2087–2094, 2006.

[6] B. Ben Amor, J. Su, and A. Srivastava. Action recognition using rate-invariant analysis of skeletal shape trajectories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(1):1–13, 2016.

[7] S. Bonnabel and R. Sepulchre. Riemannian metric and geometric mean for positive semidefinite matrices of fixed rank. *SIAM Journal on Matrix Analysis and Applications*, 31(3):1055–1070, 2009.

[8] C. A. Corneanu, M. Oliu, J. F. Cohn, and S. Escalera. Survey on rgb, 3D, thermal, and multimodal approaches for facial expression recognition: history, trends, and affect-related applications. 2016.

[9] 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 Trans. Cybernetics*, 45(7):1340–1352, 2015.

[10] M. M. Deza and M. Laurent. *Geometry of cuts and metrics*, volume 15. Springer, 2009.

[11] A. Dhall, R. Goecke, J. Joshi, M. Wagner, and T. Gedeon. Emotion recognition in the wild challenge (emotiw) challenge and workshop summary. In *2013 International Conference on Multimodal Interaction, ICMI '13, Sydney, NSW, Australia, December 9-13, 2013*, pages 371–372, 2013.

[12] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon. Collecting large, richly annotated facial-expression databases from movies. *IEEE MultiMedia*, 19(3):34–41, 2012.

[13] H. Ding, S. K. Zhou, and R. Chellappa. Facenet2expnet: Regularizing a deep face recognition net for expression recognition. *CoRR*, abs/1609.06591, 2016.

[14] S. Elaiwat, M. Bennamoun, and F. Boussaïd. A spatio-temporal rbm-based model for facial expression recognition. *Pattern Recognition*, 49:152–161, 2016.

[15] M. Faraki, M. T. Harandi, and F. Porikli. Image set classification by symmetric positive semi-definite matrices. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*, pages 1–8. IEEE, 2016.

[16] T. Graepel, R. Herbrich, P. Bollmann-Sdorra, and K. Obermayer. Classification on pairwise proximity data. *Advances in neural information processing systems*, pages 438–444, 1999.

[17] S. Gudmundsson, T. P. Runarsson, and S. Sigurdsson. Support vector machines and dynamic time warping for time series. In *Neural Networks, 2008. IJCNN 2008.(IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on*, pages 2772–2776. IEEE, 2008.

[18] S. Jain, C. Hu, and J. K. Aggarwal. Facial expression recognition with temporal modeling of shapes. In *IEEE International Conference on Computer Vision Workshops, ICCV 2011 Workshops, Barcelona, Spain, November 6-13, 2011*, pages 1642–1649, 2011.

[19] S. Jayasumana, R. I. Hartley, M. Salzmann, H. Li, and M. T. Harandi. Kernel methods on riemannian manifolds with gaussian RBF kernels. *IEEE Trans. Pattern Anal. Mach. Intell.*, 37(12):2464–2477, 2015.

[20] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim. Joint fine-tuning in deep neural networks for facial expression recognition. In *IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2983–2991, 2015.

[21] A. Kläser, M. Marszalek, and C. Schmid. A spatio-temporal descriptor based on 3d-gradients. In *Proceedings of the British Machine Vision Conference 2008, Leeds, September 2008*, pages 1–10, 2008.

[22] S. Kobayashi and K. Nomizu. *Foundations of Differential Geometry*, volume 1. Interscience Publishers, 1963.

[23] J. Kossaifi, G. Tzimiropoulos, S. Todorovic, and M. Pantic. AFEW-VA database for valence and arousal estimation in-the-wild. *Image and Vision Computing*, 2017.

[24] M. Liu, S. Li, S. Shan, R. Wang, and X. Chen. Deeply learning deformable facial action parts model for dynamic expression analysis. In *Computer Vision - ACCV 2014 - 12th Asian Conference on Computer Vision, Singapore, Singapore, November 1-5, 2014, Revised Selected Papers, Part IV*, pages 143–157, 2014.

[25] M. Liu, S. Shan, R. Wang, and X. Chen. Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pages 1749–1756, 2014.

[26] P. Lucey, J. F. Cohn, T. Kanade, J. M. Saragih, Z. Ambadar, and I. A. Matthews. The extended cohn-kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2010, San Francisco, CA, USA, 13-18 June, 2010*, pages 94–101, 2010.

[27] G. Meyer, S. Bonnabel, and R. Sepulchre. Regression on fixed-rank positive semidefinite matrices: a Riemannian approach. *Journal of Machine Learning Research*, 12(Feb):593–625, 2011.

[28] X. Pennec, P. Fillard, and N. Ayache. A riemannian framework for tensor computing. *International Journal of Computer Vision*, 66(1):41–66, 2006.

[29] R. W. Ptucha, G. Tsagkatakis, and A. E. Savakis. Manifold based sparse representation for robust expression recognition without neutral subtraction. In *IEEE International Conference on Computer Vision Workshops, ICCV 2011 Workshops, Barcelona, Spain, November 6-13, 2011*, pages 2136–2143, 2011.

[30] A. Sanin, C. Sanderson, M. T. Harandi, and B. C. Lovell. Spatio-temporal covariance descriptors for action and gesture recognition. In *2013 IEEE Workshop on Applications of Computer Vision, WACV 2013, Clearwater Beach, FL, USA, January 15-17, 2013*, pages 103–110, 2013.

[31] E. Sariyanidi, H. Gunes, and A. Cavallaro. Learning bases of activity for facial expression recognition. *IEEE Transactions on Image Processing*, PP(99):1–1, 2017.

[32] B. Scholkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA, 2001.

[33] P. Scovanner, S. Ali, and M. Shah. A 3-dimensional sift descriptor and its application to action recognition. In *Proceedings of the 15th International Conference on Multimedia 2007, Augsburg, Germany, September 24-29, 2007*, pages 357–360, 2007.

[34] J. Shen, S. Zafeiriou, G. G. Chrysos, J. Kossaifi, G. Tzimiropoulos, and M. Pantic. The first facial landmark tracking in-the-wild challenge: Benchmark and results. In *2015 IEEE International Conference on Computer Vision Workshop, ICCV Workshops 2015, Santiago, Chile, December 7-13, 2015*, pages 1003–1011, 2015.

[35] S. Taheri, P. Turaga, and R. Chellappa. Towards view-invariant expression analysis using analytic shape manifolds. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 306–313. IEEE, 2011.

[36] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1701–1708, 2014.

[37] M. F. Valstar and M. Pantic. Induced disgust, happiness and surprise: an addition to the mmi facial expression database. In *Proceedings of Int'l Conf. Language Resources and Evaluation, Workshop on EMOTION*, pages 65–70, Malta, May 2010.

[38] B. Vandereycken, P.-A. Absil, and S. Vandewalle. Embedded geometry of the set of symmetric positive semidefinite matrices of fixed rank. In *Statistical Signal Processing, 2009. SSP'09. IEEE/SP 15th Workshop on*, pages 389–392. IEEE, 2009.

[39] R. Vemulapalli, F. Arrate, and R. Chellappa. Human action recognition by representing 3d skeletons as points in a lie group. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 588–595, 2014.

[40] R. Vemulapalli and R. Chellapa. Rolling rotations for recognizing human actions from 3d skeletal data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4471–4479, 2016.

[41] L. Wang, Y. Qiao, and X. Tang. Motionlets: Mid-level 3d parts for human motion recognition. In *2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, June 23-28, 2013*, pages 2674–2681, 2013.

[42] R. Wang, H. Guo, L. S. Davis, and Q. Dai. Covariance discriminative learning: A natural and efficient approach to image set classification. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2496–2503. IEEE, 2012.

[43] Z. Wang, S. Wang, and Q. Ji. Capturing complex spatio-temporal relations among facial muscles for facial expression recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3422–3429, 2013.

[44] X. Xiong and F. D. la Torre. Supervised descent method and its applications to face alignment. In *2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, June 23-28, 2013*, pages 532–539, 2013.

[45] G. Zhao, X. Huang, M. Taini, S. Z. Li, and M. Pietikäinen. Facial expression recognition from near-infrared videos. *Image Vision Comput.*, 29(9):607–619, 2011.

[46] G. Zhao and M. Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6), 2007.

[47] G. Zhao and M. Pietikäinen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):915–928, 2007.

[48] L. Zhong, Q. Liu, P. Yang, B. Liu, J. Huang, and D. N. Metaxas. Learning active facial patches for expression analysis. In *2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012*, pages 2562–2569, 2012.