Barycentric Representation and Metric Learning for Facial Expression Recognition

Anis Kacem, Mohamed Daoudi, Juan-Carlos Alvarez-Paiva

Barycentric Representation and Metric Learning for Facial Expression Recognition

Anis Kacem[1], Mohamed Daoudi[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 tackle the problem of dynamic facial expression recognition. An affine-invariant facial shape representation based on barycentric coordinates is proposed and related to the Grassmannian representation. Unlike the latter, the barycentric representation allows us to work directly on Euclidean space and apply a metric learning algorithm to find a suitable metric that is discriminative enough to compare facial shapes under different expressions. Finally, we exploit the learned metric in a machinery combining a Dynamic Time Warping (DTW) phase and a pairwise proximity function SVM classifier for a rate-invariant classification of the facial sequences. Experiments on the AFEW dataset show the effectiveness of our approach while exploiting only geometric features.**

I. INTRODUCTION

In recent years, facial expression recognition has become a popular research field due to its wide applications in many areas such as biometrics, psychological analysis, human-computer interaction, and so on. Facial expressions involve the movements of some facial muscles but occur along with head motions and pose variations. Therefore, it is necessary for facial expression analysis to be invariant with respect to head pose changes. This is a challenging task especially due to large variations in the appearance of facial expressions from different views. Recent advances in face landmark tracking [3], [20], opened a gate to landmark-based facial expression analysis even in uncontrolled conditions [23]. However, these landmarks may be distorted by undesirable projective transformations accentuated by head pose changes. These projective transformations can be approximated by affine transformations, especially when the face is far from the camera [29]. Hence, filtering out the affine transformations is a convenient way to handle head pose changes. Accordingly, some works encoded the landmarks in the Grassmannian, guaranteeing the affine invariance [5], [29], [21]. However, this representation leads to the usual difficulties in the handling of *nonlinear data*. To overcome this issue, many sophisticated learning methods [19], [29], [30], [31], [18], [1], [21] have been devised to linearize the data while respecting the geometry of the Grassmann manifold. 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. Some authors propose to embed a manifold in a high dimensional Reproducing Kernel Hilbert Space (RKHS), by exploiting a positive definite kernel function to embed the manifold into a reproducing kernel Hilbert space [19]. In another context, a projection metric learning method on the Grassmannian was proposed in [17]. More recently, a deep architecture that performs deep learning over Grassmann manifolds has been proposed in [18].

A more naive approach is to question the notion of nonlinear data. Manifolds of dimension n are after all nothing but $\mathbb{R}^n$ with lower-dimensional pieces or cells glued in. If the data falls outside of lower dimensional pieces, we may in principle consider the data as linear. In the case at hand, the points on the Grassmannian corresponding to the facial landmarks are naturally contained in one of the standard charts. It turns out that passing to this chart is nothing more than taking barycentric coordinates with respect to a specific triplet of landmark points.

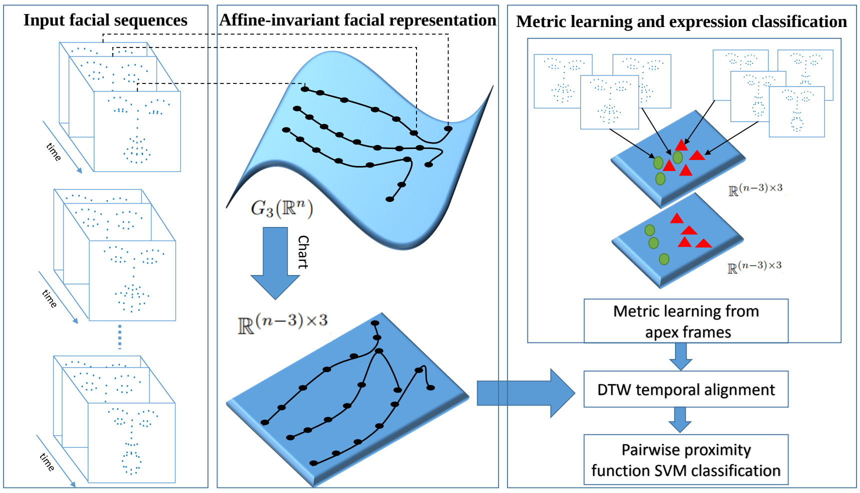
In summary, the main contributions of this paper are:

- A novel affine-invariant shape representation of the facial shapes through their barycentric coordinates, resulting in vectors lying in Euclidean space. We show how this representation is related to the conventional Grassmann representation;
- A metric learning algorithm is applied to find a suitable metric for a more discriminative comparison of facial shapes;
- A rate-invariant similarity based learning process combining a Dynamic Time Warping equipped with the learned metric, and a pairwise proximity function SVM (ppfSVM) classifier for expression recognition.

Fig. 1 shows an overview of the proposed approach. The rest of the paper is organized as following. In section II, we propose a new barycentric representation of facial shape. The barycentric representation allows us to work directly on Euclidean space and apply a metric learning algorithm to find a suitable metric that is discriminative enough to compare facial shapes under different expressions. Section III states the classification approach. Experimental results and discussions are reported in section IV. In section V, we conclude and draw some perspectives of the work.

II. FACIAL SHAPE REPRESENTATION

A basic mathematical problem that arises in facial shape analysis is to study the motion of an ordered list of land-



Input facial sequences
time
time
time
Affine-invariant facial representation
$G_3(\mathbb{R}^n)$
Chart
$\mathbb{R}^{(n-3)\times 3}$
Metric learning and expression classification
$\mathbb{R}^{(n-3)\times 3}$
$\mathbb{R}^{(n-3)\times 3}$
Metric learning from apex frames
DTW temporal alignment
Pairwise proximity function SVM classification

Fig. 1. Overview of the proposed approach – After automatic landmark detection for each frame of the video, we represent the resulting shapes through their barycentric coordinates. While being closely related to the affine-invariant Grassmann representation, this representation allows us to work directly on Euclidean space where a metric learning algorithm is applied. Dynamic Time Warping (DTW) using the learned metric is then performed to align the facial sequences. Finally, the ppfSVM exploiting the DTW similarity measure is used as expression classifier.

marks, $Z_1(t) = (x_1(t), y_1(t)), \ldots, Z_n(t) = (x_n(t), y_n(t))$, in the plane up to the action of an arbitrary affine transformation. A standard technique is to consider the span of the columns of the $n \times 3$ time-dependent matrix

$$M(t) := \begin{pmatrix} x_1(t) & y_1(t) & 1 \\ \vdots & \vdots & \vdots \\ x_n(t) & y_n(t) & 1 \end{pmatrix} .$$

If for every time t there exists some triplet of landmarks forming a non-degenerate triangle the rank of the matrix $M(t)$ is constantly equal to 3 and the span of its columns is a curve of three-dimensional subspaces in $\mathbb{R}^n$. In other words, a curve in the Grassmannian $G_3(\mathbb{R}^n)$, which is well known [5], [29], [21] to be an affine invariant of the motion. This convenient way of filtering out the affine transformations opens the way to the use of metric and differential-geometric techniques in the study and classification of moving landmarks [30], [6], [10], [21], [2].

Another convenient and more classic way to filter out affine transformations is through the use of *barycentric coordinates*. This method can be applied provided three of the landmarks form a non-degenerate triangle throughout all their motion. Indeed, assume, without loss of generality, that $Z_1(t)$, $Z_2(t)$, and $Z_3(t)$ are the vertices of a non-degenerate triangle *for every value of* t. For every number $i = 4, \ldots, n$ and every time t we can write

$$Z_i(t) = \lambda_{i1}(t) Z_1(t) + \lambda_{i2}(t) Z_2(t) + \lambda_{i3}(t) Z_3(t) \ ,$$

where the numbers $\lambda_{i1}(t)$, $\lambda_{i2}(t)$, and $\lambda_{i3}(t)$ satisfy

$$\lambda_{i1}(t) + \lambda_{i2}(t) + \lambda_{i3}(t) = 1.$$

This last condition renders the triplet of barycentric coordinates $(\lambda_{i1}(t), \lambda_{i2}(t), \lambda_{i3}(t))$ unique. In fact, it is equal to

$$(x_i(t), y_i(t), 1) \begin{pmatrix} x_1(t) & y_1(t) & 1 \\ x_2(t) & y_2(t) & 1 \\ x_3(t) & y_3(t) & 1 \end{pmatrix}^{-1} .$$

If T is an affine transformation of the plane, the barycentric representation of $TZ_i(t)$ in terms of the frame given by $TZ_1(t)$, $TZ_2(t)$, and $TZ_3(t)$ is still $(\lambda_{i1}(t), \lambda_{i2}(t), \lambda_{i3}(t))$. This allows us to propose the $(n-3) \times 3$ matrix

$$\Lambda(t) := \begin{pmatrix} \lambda_{41}(t) & \lambda_{42}(t) & \lambda_{43}(t) \\ \vdots & \vdots & \vdots \\ \lambda_{n1}(t) & \lambda_{n2}(t) & \lambda_{n3}(t) \end{pmatrix} .$$

as the affine shape representation of the moving landmarks.

A. Barycentric Representation and Grassman Representation

In order to expose the basic relationship between the Grassmannian representation and the barycentric representation, let us recall, in a particular case, the usual way to construct charts in the Grassmannian. If $\zeta \in G_3(\mathbb{R}^n)$ is a subspace that intersects the $(n-3)$-dimensional subspace

$$W = \{(0, 0, 0, x_4, \ldots, x_n) : x_i \in \mathbb{R}^n \text{ for } i \text{ between } 4 \text{ and } n\}$$

only at the origin, and $\mathbf{x} = (x_1, \ldots, x_n)$, $\mathbf{y} = (y_1, \ldots, y_n)$, and $\mathbf{z} = (z_1, \ldots, z_n)$ is a basis for ζ, then the 3×3 matrix

$$\begin{pmatrix} x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \\ x_3 & y_3 & z_3 \end{pmatrix}$$

is invertible and the $(n-3) \times 3$ matrix

$$\begin{pmatrix} x_4 & y_4 & z_4 \\ \vdots & \vdots & \vdots \\ x_n & y_n & z_n \end{pmatrix} \begin{pmatrix} x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \\ x_3 & y_3 & z_3 \end{pmatrix}^{-1}$$

is independent of the chosen basis. In this way, the open and dense set of 3-dimensional subspaces transverse to W are put in a bijective correspondence with $\mathbb{R}^{(n-3)\times 3}$.

If we consider the curve in $G_3(\mathbb{R}^n)$ given by the span of the columns of the matrix

$$M(t) := \begin{pmatrix} x_1(t) & y_1(t) & 1 \\ \vdots & \vdots & \vdots \\ x_n(t) & y_n(t) & 1 \end{pmatrix}$$

and if the landmarks $Z_1(t) = (x_1(t), y_1(t))$, $Z_2(t) = (x_2(t), y_2(t))$, and $Z_3(t) = (x_3(t), y_3(t))$ form a non-degenerate triangle throughout all their motion, then composing this curve with chart in the Grassmannian yields the curve of matrices

$$\begin{pmatrix} x_4(t) & y_4(t) & 1 \\ \vdots & \vdots & \vdots \\ x_n(t) & y_n(t) & 1 \end{pmatrix} \begin{pmatrix} x_1(t) & y_1(t) & 1 \\ x_2(t) & y_2(t) & 1 \\ x_3(t) & y_3(t) & 1 \end{pmatrix}^{-1} ,$$

which is just the curve $\Lambda(t)$ encoding the barycentric representation of the landmarks. For more details about the affine-invariance with barycentric coordinates, please refer to the page 81 of the book [7]. In what follows, we will consider the introduced affine-invariant vector Λ, with dimension $m = (n-3) \times 3$, to represent a static facial shape and the curve $\Lambda(t)$ to denote a facial shape sequence.

B. Metric learning of Affine Shape Representation

Given the facial shape represented by the affine-invariant vector Λ, with dimension $m = (n-3)\times 3$, we seek a suitable metric that is discriminative enough in terms of expression to compare them. The Euclidean distance, defined as the squared l_2-norm of the difference of the vectors, could be a reasonable choice since the defined shapes lie in Euclidean space. However, such distance disregards the specific nature of the considered facial shapes. To overcome this issue, we propose to learn a Mahalanobis distance instead of using the standard Euclidean distance [24]. Given two facial shapes represented by the affine-invariant vectors Λ_i and Λ_j in $\mathbb{R}^m$, the Mahalanobis distance is defined by

$$d^2_{l_{ij}}(\Lambda_i, \Lambda_j) = (\Lambda_i - \Lambda_j)^T A (\Lambda_i - \Lambda_j) \ , \qquad (1)$$

where A is a positive semidefinte (p.s.d) matrix of size $m \times m$. The problem of metric learning is then to find the best p.s.d matrix A that best discriminates the facial expressions, *i.e.,* results in small distances when the facial shapes represent similar expressions and large distances when they represent different expressions.

Let $\mathcal{D} = \{(\Lambda_1, c_1), \ldots, (\Lambda_N, c_N)\}$ represent a set of affine-invariant shapes in $\mathbb{R}^m$ annotated with the corresponding expressions (*e.g.*, $c =$'happy', 'angry', etc.). Let $\{\Lambda_i, \Lambda_j, \Lambda_k\}$ be a triplet of affine-invariant shapes from $\mathcal{D}$ such that (Λ_i, Λ_j) have the same label ($c_i = c_j$), and (Λ_i, Λ_k) with different labels ($c_i \neq c_k$). We aim to find an optimal p.s.d matrix A such that $d^2_{l_{ij}}(\Lambda_i, \Lambda_j) < d^2_{l_{ik}}(\Lambda_i, \Lambda_k)$. That is, we wish to find a p.s.d matrix A that minimizes $d^2_{l_{ij}} - d^2_{l_{ik}} = (\Lambda_i - \Lambda_j)^T A (\Lambda_i - \Lambda_j) - (\Lambda_i - \Lambda_k)^T A (\Lambda_i - \Lambda_k)$. In order to solve this optimization problem, we follow the convenient method described by Shen *et al.* [28], where a boosting is used.

III. FACIAL SEQUENCE CLASSIFICATION

The learned distance does, indeed, assign small distances to similar static facial shapes and large distances to dissimilar shapes. However, as conveying an expression is a temporal process, we are more interested in comparing facial shape sequences. Accordingly, we exploit the learned distance to build a rate-invariant similarity measure between facial shape sequences. Specifically, the Dynamic Time Warping (DTW) algorithm [8], employing the learned distance instead of the standard Euclidean distance, is used to compare two facial sequences.

Following [4], [21], we adopt the *pairwise proximity function SVM* (ppfSVM) [14], [15] to classify the facial sequences. PpfSVM requires the definition of a similarity measure to compare samples. In our case, it is natural to consider the similarity measure given by our version of DTW for such a comparison. This strategy involves the construction of inputs such that each sequence is represented by its similarity to all the sequences in the dataset, with respect to the DTW similarity measure, and then apply a conventional SVM to this transformed data [15]. The ppfSVM is related to the arbitrary kernel-SVM without restrictions on the kernel function [14]. Further details on ppfSVM can be found in [4], [14], [15].

IV. EXPERIMENTAL RESULTS

A. Experimental settings

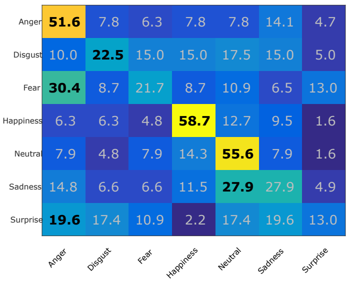
In order to learn the metric, we use only peak frames from each facial sequence, where the expression reaches its peak. Since peak frames are difficult to detect in spontaneous facial expressions, we performed the metric learning using extracted landmarks from CK+ dataset [26] which is captured in strict controlled conditions. In this dataset, 309 facial sequences of 118 subjects are annotated with the six labels (the six basic emotions). In all the sequences, the actors start by being neutral then perform the expression until reaching a peak. In our experiments, we only used the five last frames and the first frame from all the sequences. The labels of the five last frames are assigned according to the label of the sequence, while the label of the first frame is always considered as 'neutral'. A total number of 16686 facial shapes are used for the training phase to learn the Mahalanobis distance.

To evaluate the proposed approach, we conducted experiments on the well-known AFEW dataset [12]. This dataset contains 1106 facial sequences 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 [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'). Here we only report our results on the validation set for comparison with [11], [13], [25], [16]. Note that our experiments are made once the facial landmarks are extracted using the method proposed in [3]. The three points used to form the non-degenerate triangle, essential to build the affine-invariant shapes from the landmarks, are the points positioned at the left and right corners of the eye and the nose tip.

All our programs were implemented in Matlab and run on a 2.8 GHZ CPU. We used the multi-class SVM implementation of the LibSVM library [9], and the codes given by [28] for the metric learning.

B. Results and discussions

Following the experimental settings mentioned in the previous section, we report an accuracy of 38.38%. From the corresponding confusion matrix shown in Fig. 2, we can observe that the highest performances are obtained for 'Anger' (51.6%), 'Happiness' (58.7%), and 'Neutral' (55.6%). Since AFEW is a very challenging dataset, the obtained results are competitive with state-of-art approaches as shown in Table I. We recorded better performance than many appearance based approaches such as SPDNet [16] and STM-ExpLet [25]. However, our results are outperformed by [21] where Gram matrices are used to represent facial shapes and compared with a defined Riemannian metric. The execution time of comparing two arbitrary sequences on AFEW dataset is 0.064 seconds with the proposed approach against 0.84 seconds with the approach proposed in [21]. In Table I, we can observe that our results compared to [21] are outperformed by only 1% while being 10 times faster.



	Anger	Disgust	Fear	Happiness	Neutral	Sadness	Surprise
Anger	**51.6**	7.8	6.3	7.8	7.8	14.1	4.7
Disgust	10.0	**22.5**	15.0	15.0	17.5	15.0	5.0
Fear	**30.4**	8.7	21.7	8.7	10.9	6.5	13.0
Happiness	6.3	6.3	4.8	**58.7**	12.7	9.5	1.6
Neutral	7.9	4.8	7.9	14.3	**55.6**	7.9	1.6
Sadness	14.8	6.6	6.6	11.5	**27.9**	27.9	4.9
Surprise	**19.6**	17.4	10.9	2.2	17.4	19.6	13.0

Fig. 2. Confusion matrix on AFEW dataset

TABLE I
OVERALL ACCURACY AFEW DATASET

Method	**Accuracy** (%)
$^{(A)}$ HOG 3D [22]	26.90
$^{(A)}$ HOE [32]	19.54
$^{(A)}$ 3D SIFT [27]	24.87
$^{(A)}$ LBP-TOP [33]	25.13
$^{(A)}$ EmotiW [11]	27.27
$^{(A)}$ STM [25]	29.19
$^{(A)}$ STM-ExpLet [25]	31.73
$^{(A)}$ SPDNet [16]	34.23
$^{(G)}$ Gram Trajectories [21]	**39.94**
$^{(G)}$ **Ours**	**38.38**

To evaluate the different steps of the proposed pipeline, we performed baseline experiments. Firstly, we conducted the same experiments while using alternative representations and metrics. We compared our results with a conventional Grassmann affine-invariant representation coupled with a Riemannian metric given by the subspace angles [5], [29], [21]. The achieved accuracy is around 2.5% lower than ours. We also replaced the learned Mahalanobis distance with a standard Euclidean distance. Here also, the performance decreases by about 3%. In Table II, we show the achieved accuracies by the described alternative representations and metrics and the necessary execution time to compare two arbitrary facial shapes. One can observe that the proposed representation achieves better performance than the Grassmannian while being less time consuming. These results show the effectiveness of the proposed representation and the importance of the metric learning step in our pipeline. As mentioned in the previous section, we used the five last (peak) frames from the sequences of CK+ dataset to learn the Mahalanobis distance. In Table II, we provide the obtained accuracies when using one, two, five and seven last peak frames from each sequence. The highest accuracy is obtained with the last five frames. Besides, we report in Table II the average accuracy when DTW is used or not in our pipeline. It is clear from these experiments that a temporal alignment is an important step as an improvement of around 7% is obtained . In the last table, we compare the results of ppfSVM to a K-NN classifier coupled with the DTW similarity measure. The number of nearest neighbors K is chosen by cross-validation. We obtained an average accuracy of 31.33% for $K = 5$. These results are outperformed by ppfSVM classifier.

TABLE II
BASELINE EXPERIMENTS

Distance	**Accuracy** (%)	**Time** (μs)
Subspace angles in $G_3(\mathbb{R}^n)$	36.81	2967
Euclidean distance	36.55	530
Mahalanobis distance d_l	**38.38**	**568**

Number of peak frames	**Accuracy** (%)
1 peak frame	37.07
2 peak frames	37.59
5 peak frames	**38.38**
7 peak frames	36.29

Temporal alignment	**Accuracy** (%)	**Time** (s)
without DTW	30.8	0.008
with DTW	**38.38**	**0.064**

Classifier	**Accuracy** (%)
K-NN	31.33
ppf-SVM	**38.38**

V. CONCLUSIONS AND FUTURE WORKS

In this paper, we proposed an affine-invariant facial shape through encoding the corresponding landmark points by their barycentric coordinates. Such representation results in vectors lying in an Euclidean space where a multitude of Euclidean metrics is applicable. A Finally, a learned metric is incorporated in the DTW similarity measure. The first experimental results on AFEW dataset showing the effectiveness of the affine-shape representation, encourage us to improve the metric learning and the classification pipeline.

References

[1] T. Alashkar, B. Ben Amor, M. Daoudi, and S. Berretti. Spontaneous expression detection from 3D dynamic sequences by analyzing trajectories on Grassmann manifolds. *IEEE Transactions on Affective Computing*, 2018. To appear.

[2] R. Anirudh, P. K. Turaga, J. Su, and A. Srivastava. Elastic functional coding of Riemannian trajectories. *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] M. Berger. Geometry, vol. i-ii, 1987.

[8] D. J. Berndt and J. Clifford. Using dynamic time warping to find patterns in time series. In *KDD workshop*, volume 10, pages 359–370. Seattle, WA, 1994.

[9] C.-C. Chang and C.-J. Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27, 2011.

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

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

[14] 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.

[15] S. Gudmundsson, T. P. Runarsson, and S. Sigurdsson. Support vector machines and dynamic time warping for time series. In *IEEE International Joint Conference on Neural Networks*, pages 2772–2776. IEEE, 2008.

[16] Z. Huang and L. J. Van Gool. A Riemannian network for spd matrix learning. In *AAAI*, volume 2, page 6, 2017.

[17] Z. Huang, R. Wang, S. Shan, and X. Chen. Projection metric learning on Grassmann manifold with application to video based face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 140–149, 2015.

[18] Z. Huang, J. Wu, and L. V. Gool. Building deep networks on Grassmann manifolds. *CoRR*, abs/1611.05742, 2016.

[19] S. Jayasumana, R. I. Hartley, M. Salzmann, H. Li, and M. T. Harandi. Kernel methods on Riemannian manifolds with gaussian RBF kernels. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(12):2464–2477, 2015.

[20] A. Jourabloo and X. Liu. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4188–4196, 2016.

[21] A. Kacem, M. Daoudi, B. Ben Amor, and J. C. Alvarez-Paiva. A novel space-time representation on the positive semidefinite cone for facial expression recognition. In *IEEE International Conference on Computer Vision (ICCV)*, October 2017.

[22] 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.

[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] B. Kulis. Metric learning: A survey. *Foundations and Trends in Machine Learning*, 5(4):287–364, 2013.

[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] 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.

[28] C. Shen, J. Kim, L. Wang, and A. Hengel. Positive semidefinite metric learning with boosting. In *Advances in neural information processing systems*, pages 1651–1659, 2009.

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

[30] 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.

[31] 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.

[32] 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.

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