Detecting Depression Severity by Interpretable Representations of Motion Dynamics

Anis Kacem, Zakia Hammal, Mohamed Daoudi, Jeffrey Cohn

Detecting Depression Severity by Interpretable Representations of Motion Dynamics

Anis Kacem[1], Zakia Hammal[2], Mohamed Daoudi[1], and Jeffrey Cohn[2,3]
[1] IMT Lille Douai, Univ. Lille, CNRS, UMR 9189 CRIStAL, F-59000 Lille, France
[2] Robotics Institute, Carnegie Mellon University, USA
[3] Department of Psychology, University of Pittsburgh, USA

***Abstract*— Recent breakthroughs in deep learning using automated measurement of face and head motion have made possible the first objective measurement of depression severity. While powerful, deep learning approaches lack interpretability. We developed an interpretable method of automatically measuring depression severity that uses barycentric coordinates of facial landmarks and a Lie-algebra based rotation matrix of 3D head motion. Using these representations, kinematic features are extracted, preprocessed, and encoded using Gaussian Mixture Models (GMM) and Fisher vector encoding. A multi-class SVM is used to classify the encoded facial and head movement dynamics into three levels of depression severity. The proposed approach was evaluated in adults with history of chronic depression. The method approached the classification accuracy of state-of-the-art deep learning while enabling clinically and theoretically relevant findings. The velocity and acceleration of facial movement strongly mapped onto depression severity symptoms consistent with clinical data and theory.**

I. INTRODUCTION

Many of the symptoms of depression are observable. In depression facial expressiveness [23], [26] and head movement [12], [18], [14] are reduced. The velocity of head movement also is slower in depression [14].

Yet, systematic means of using observable behavior to inform screening and diagnosis of the occurrence and severity of depression are lacking. Recent advances in computer vision and machine learning have explored the validity of automatic measurement of depression severity from video sequences [1], [28], [31], [8].

Hdibeklioglu and colleagues [8] proposed a multimodal deep learning based approach to detect depression severity in participants undergoing treatment for depression. Deep learning based per-frame coding and per-video Fisher-vector based coding were used to characterize the dynamics of facial and head movement. For each modality, selection among features was performed using combined mutual information, which improved accuracy relative to blanket selection of all features regardless of their merit. For individual modalities, facial and head movement dynamics outperformed vocal prosody. For combinations, fusing the dynamics of facial and head movement was more discriminative than head movement dynamics and more discriminative than facial movement dynamics plus vocal prosody and head movement dynamics plus vocal prosody. The proposed deep learning based method outperformed the state of the art counterparts for each modality.

A limitation of the deep learning approach is its lack of interpretability. The dynamics of facial, head, and vocal prosody were important, but the nature of those changes during course of depression were occult. From their findings, one could not say whether dynamics were increasing, decreasing, or varying in some non-linear way. For clinical scientists and clinicians interested in the mechanisms and course of depression, interpretable features matter. They want to know not only presence or severity of depression but how dynamics vary with occurrence and severity of depression.

Two previous shallow-learning approaches to depression detection were interpretable but less sensitive to depression severity. In Alghowinem and colleagues [1], head movements were tracked by AAMs [25] and modeled by Gaussian mixture models with seven components. Mean, variance, and component weights of the learned GMMs were used as features. And a set of interpretable head pose functionals was proposed. These included the statistics of head movements and duration of looking in different directions.

Williamson and his colleagues [31] investigated the specific changes in coordination, movement, and timing of facial and vocal signals as potential symptoms for self-reported BDI (Beck Depression Inventory) scores [2]. They proposed a multi-scale correlation structure and timing feature sets from video-based facial action units (AUs [10]) and audio-based vocal features. The features were combined using a Gaussian mixture model and extreme learning machine classifiers to predict BDI scores.

Reduced facial expression is commonly observed in depression and relates to deficits in experiencing positive as well as negative emotion [24]. Less often, greatly increased expression occurs. There are referred to as psychomotor retardation and psychomotor agitation, respectively. We propose to capture aspects of psychomotor retardation and agitation using the dynamics of facial and head movement. Participants were from a clinical trial for treatment of moderate to severe depression and had history of multiple depressive episodes. Compared to state-of-the-art deep learning approch for depression severity assessment, we propose a reliable and clinically interpretable method of automatically measuring depression severity from the dynamics of face and head motion.

To analyze facial movement dynamics separately from head movement dynamics, facial shape representation would need to be robust to head pose changes while preserving

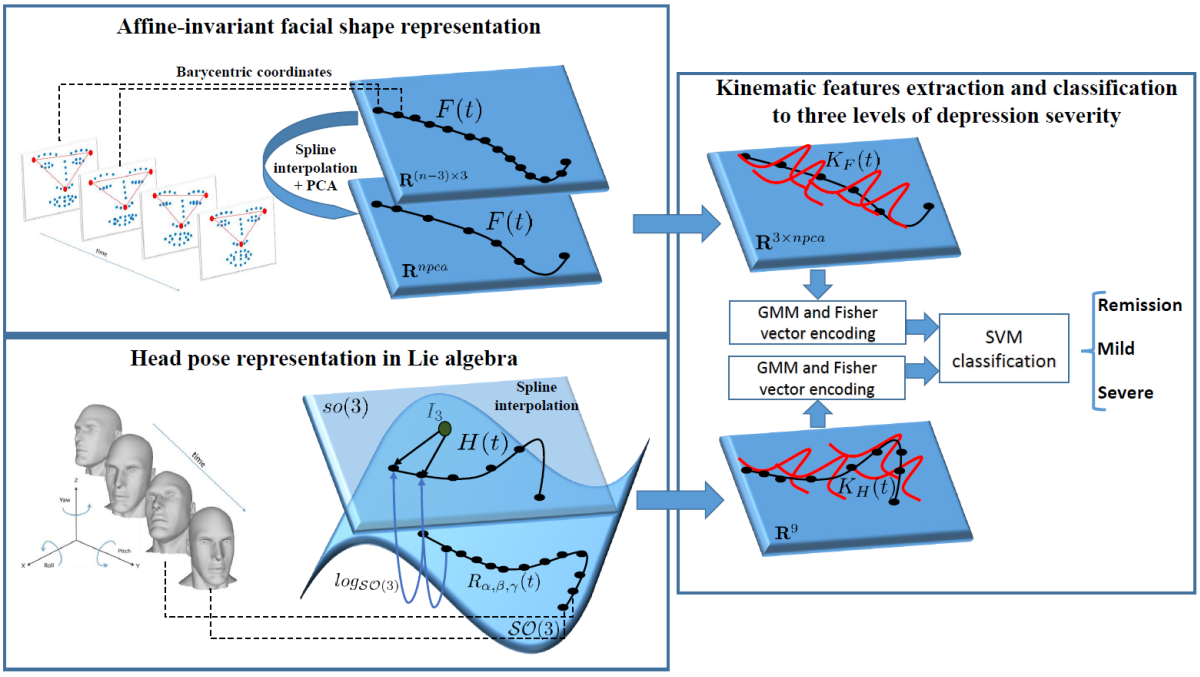

Fig. 1. Overview of the proposed approach.

facial motion information [3], [27], [19], [16]. To achieve this goal, Kacem and colleagues [20] used the Gram matrix of facial landmarks to obtain a facial representation that is invariant to Euclidean transformations (*i.e.,* rotations and translations). In related work, Begel and colleagues [3] and Taheri and colleagues [27] used mapping facial landmarks in Grassmann manifold to achieve an affine-invariance representation. These previous efforts yielded to facial shape representations lying on non-linear manifolds where standard Euclidean analysis techniques are not straightforward. Based on our work [19], we propose an efficient representation for facial shapes through encoding landmark points by their *barycentric coordinates* [4]. In addition to the affine-invariance, the proposed approach has the advantage of lying on Euclidean space avoiding the non-linearity problem.

Because we are interested in both facial movement dynamics and head movement dynamics, the later is encoded by combining the 3 degrees of freedom of head movement (*i.e.,* yaw, roll, and pitch angles) in a single rotation matrix mapped to Lie algebra to overcome the non-linearity of the space of rotation matrices [29], [30].

To capture changes in the dynamics of head and facial movement that would reflect the psychomotor retardation of depressed participants, relevant kinematic features are extracted (*i.e.,* velocities and accelerations) from each proposed representation. Gaussian Mixture Models (GMM) combined with an improved fisher vector encoding are then used to obtain a single vector representation for each sequence (*i.e.,* interview). Finally, a multi-class SVM with a Gaussian kernel is used to classify the encoded facial and head movement dynamics into three depression severity levels. The overview of the proposed approach is shown in Fig. I.

The main contributions of this paper are three:

- An affine-invariant facial shape representation that is robust to head pose changes through encoding the landmark points by their barycentric coordinates.
- A natural head pose representation in Lie algebra with respect to the geometry of the space of head rotations.
- Extraction and classification of kinematic features that encode well the dynamics of facial and head movements for the purpose of depression severity level assessment and are interpretable and consistent with data and theory in depression.

The rest of the paper is organized as follows: In section II facial shape representation and head pose representation are presented. In section III, we describe kinematic features based on the representations proposed in section II. Section IV describes the depression severity level classification approach. Results and discussions are reported in section V. In section VI, we conclude and draw some perspectives of the work.

II. FACIAL SHAPE AND HEAD POSE REPRESENTATION

We propose an automatic and interpretable approach for the analysis of facial and head movement dynamics for depression severity assessment.

A. Automatic Tracking of Facial Landmarks and Head Pose

Zface [17], an automatic, person-independent, generic approach was used to track the 2D coordinates of 49 facial landmarks (fiducial points) and 3 degrees of out-of-plane rigid head movements (*i.e.,* pitch, yaw, and roll) from 2D videos. Because our interest is the dynamics rather than

the configuration we used the facial and head movement dynamics for the assessment of depression severity. Facial movement dynamics is represented using the time series of the coordinates of the 49 tracked fiducial points. Likewise, head movement dynamics is represented using the time series of the 3 degrees of freedom of out-of-plane rigid head movement.

B. Facial Shape Representation

Facial landmarks may be distorted by head pose changes that could be approximated by affine transformations. Hence, filtering out the affine transformations is a convenient way to eliminate head pose changes. In this section we briefly review the main definitions of the affine-invariance with *barycentric coordinates* and their use in facial shape analysis [19].

Our goal is to study the motion of an ordered list of landmarks, $Z_1(t) = (x_1(t), y_1(t)), \ldots, Z_n(t) = (x_n(t), y_n(t))$, 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

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

If for every time t there exists a triplet of landmarks

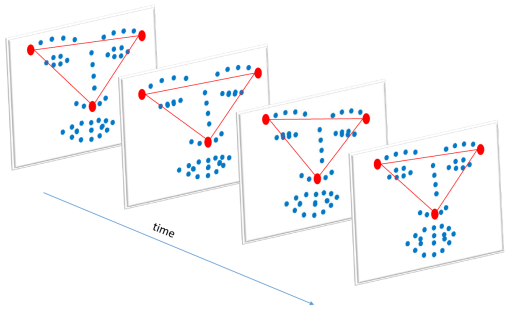

Fig. 2. Example of the automatically tracked 49 facial landmarks. The three red points denote the facial landmarks used to form the non-degenerate triangle required to compute the barycentric coordinates.

forming a non-degenerate triangle the rank of the matrix $G(t)$ is constantly equal to 3 yielding to affine-invariant representations in the non-linear Grassmann manifold of three-dimensional subspaces in $\mathbb{R}^n$. To overcome the non-linearity of the space of face representations while filtering out the affine transformations, we propose to use the *barycentric coordinates*.

Assume that $Z_1(t)$, $Z_2(t)$, and $Z_3(t)$ are the vertices of a non-degenerate triangle *for every value of* t. In the case of facial shapes, the right and left corners of the eyes and the tip of the nose are chosen to form a non-degenerate triangle (see the red triangle in Fig. 2). 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

$$F(t) := \begin{pmatrix} \lambda_{41}(t) & \lambda_{42}(t) & \lambda_{43}(t) \\ \vdots & \vdots & \vdots \\ \lambda_{n1}(t) & \lambda_{n2}(t) & \lambda_{n3}(t) \end{pmatrix} . \qquad (1)$$

as the affine invariant shape representation of the moving landmarks. It turns out that such representation is closely related to the standard Grassmannian representation while avoiding the non-linearity of the space of representations. Further details about the relationship between the barycentric and Grassmannian representations can be found in [19]. In the following, facial shape sequences are represented with the affine-invariant curve $F(t)$, with dimension $m = (n-3) \times 3$.

C. Head Pose Representation

Head movements correspond to head nods (*i.e.,* pitch), head turns (*i.e.,* yaw), and lateral head inclinations (*i.e.,* roll) (see Fig. 3). Given a time series of the 3 degrees of freedom of out-of-plane rigid head movement, for every time t the yaw is defined as a counterclockwise rotation of $\alpha(t)$ about the z-axis. The corresponding time-dependent rotation matrix is given by

$$R_\alpha(t) := \begin{pmatrix} \cos(\alpha(t)) & -\sin(\alpha(t)) & 0 \\ \sin(\alpha(t)) & \cos(\alpha(t)) & 0 \\ 0 & 0 & 1 \end{pmatrix} .$$

Pitch is a counterclockwise rotation of $\beta(t)$ about the y-axis. The rotation matrix is given by

$$R_\beta(t) := \begin{pmatrix} \cos(\beta(t)) & 0 & \sin(\beta(t)) \\ 0 & 1 & 0 \\ -\sin(\beta(t)) & 0 & \cos(\beta(t)) \end{pmatrix} .$$

Roll is a counterclockwise rotation of $\gamma(t)$ about the x-axis. The rotation matrix is given by

$$R_\gamma(t) := \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos(\gamma(t)) & -\sin(\gamma(t)) \\ 0 & \sin(\gamma(t)) & \cos(\gamma(t)) \end{pmatrix} .$$

A single rotation matrix can be formed by multiplying the yaw, pitch, and roll rotation matrices to obtain

$$R_{\alpha,\beta,\gamma}(t) = R_\alpha(t) R_\beta(t) R_\gamma(t) \ . \qquad (2)$$

The obtained time-parametrized curve $R_{\alpha,\beta,\gamma}(t)$ encodes head pose at each time t and lie on a non-linear manifold called the special orthogonal group. The special orthogonal group $\mathcal{SO}(3)$ is a matrix Lie group formed by all rotations about the origin of three-dimensional Euclidean space $\mathbb{R}^3$

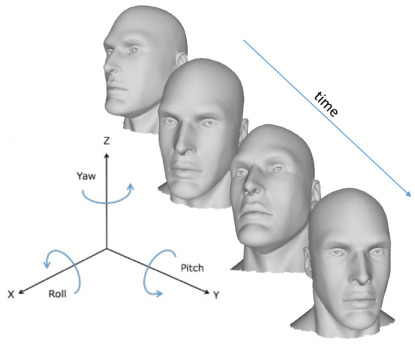

Fig. 3. Example of the automatically tracked 3 degrees of freedom of head pose.

under the operation of composition [5]. The tangent space at the identity $I_3 \in \mathcal{SO}(3)$ is a three-dimensional vector space, called the Lie algebra of $\mathcal{SO}(3)$ and is denoted by $\mathsf{so}(3)$. Following [30], [29], we overcome the non-linearity of the space of our representation (*i.e.*, $\mathcal{SO}(3)$), and map the curve $R_{\alpha,\beta,\gamma}(t)$ from $\mathcal{SO}(3)$ to $\mathsf{so}(3)$ using the logarithm map $\log_{\mathcal{SO}(3)}$ to obtain the three-dimensional curve

$$H(t) = \log_{\mathcal{SO}(3)}(I_3, R_{\alpha,\beta,\gamma}(t)) = \log(R_{\alpha,\beta,\gamma}(t)) \ , \quad (3)$$

lying on $\mathsf{so}(3)$. Fore more details about the special orthogonal group, the logarithm map, and the lie algebra, readers are refereed to [30], [29], [5]. In the following, the time series of the 3 degrees of freedom of rigid head movement are represented using the three dimensional curve $H(t)$.

III. KINEMATIC FEATURES AND FISHER VECTOR ENCODING

To characterize facial and head movement dynamics, we derive appropriate kinematic features based on their proposed representations $F(t)$ and $H(t)$, respectively.

A. Kinematic Features

Because videos of interviews varied in length, the extracted facial and head curves (of different videos) varies in length. The variation in the obtained curves' lengths may introduce distortions in the feature extraction step. To overcome this limitation, we apply a cubic spline interpolation to the obtained $F(t)$ and $H(t)$ curves, resulting in smoother, shorter, and fixed length curves. We set empirically the new length of the curve given by spline interpolation to 5000 samples for both facial and head curves.

Usually, the number of landmark points given by recent landmark detectors vary from 40 to 70 points. By building the barycentric coordinates of the facial shape as explained in section II-B, this results in high-dimensional facial curves $F(t)$ with static observations of dimension 120 at least (it can reach 200 if we have 70 landmark points per face). To reduce the dimensionality of the facial curve $F(t)$, we perform a Principal Component Analysis (PCA) that accounts for 98% of the variance to obtain new facial curves with dimension 20. Then, we compute the velocity $V_F(t) = \frac{\partial F(t)}{\partial t}$ and the acceleration $A_F(t) = \frac{\partial^2 F(t)}{\partial t^2}$ from the facial sequence $F(t)$ after reducing its dimension. Finally, facial shapes, velocities, and accelerations are concatenated to form the curve

$$K_F(t) = [F(t); V_F(t); A_F(t)] \ , \quad (4)$$

Because head curve $H(t)$ is only three-dimensional no need for data reduction. Velocities and accelerations are directly computed from the head sequence $H(t)$ and concatenated with head pose values to obtain the final nine-dimensional curve

$$K_H(t) = [H(t); V_H(t); A_H(t)] \ . \quad (5)$$

The curves $K_F(t)$ and $K_H(t)$ denote the kinematic features over time of the facial and head movements, respectively.

B. Fisher Vector Encoding

Our goal is to obtain a single vector representation from the kinematic curves $K_F(t)$ and $K_H(t)$ for depression severity assessment. Following [8], we used the Fisher Vector representation using a Gaussian mixture model (GMM) distributions [32]. Assuming that the observations of a single kinematic curve are statistically independent, a GMM with c components is computed for each kinematic curve by optimizing the maximum likelihood (ML) criterion of the observations to the c Gaussian distributions. In order to encode the estimated Gaussian distributions in a single vector representation, we use the convenient improved fisher vector encoding which is suitable for large-scale classification problems [22]. This step is performed for kinematic curves $K_F(t)$ and $K_H(t)$, separately. The number of Gaussian distributions c are chosen by a a leave-one-subject-out cross-validation and are set to 14 for kinematic facial curves and to 31 for kinematic head curves resulting in fisher vectors with dimension $14 \times 20 \times 3 \times 2 = 1680$ for facial movement dynamics and vectors with dimension $31 \times 3 \times 3 \times 2 = 558$ for head movement dynamics.

IV. ASSESSMENT OF DEPRESSION SEVERITY LEVEL

After extracting the fisher vectors from the kinematic curves, the facial and head movements are represented by compact vectors that describe the dynamics of facial and head movements, respectively. To reduce redundancy and select the most discriminative feature set, the Min-Redundancy Max-Relevance (mRMR) algorithm [21] was used for feature selection. The set of selected features are then fed to a multi-class SVM with a Gaussian kernel to classify the extracted facial and head movement dynamics into different depression severity levels. Please note that a leave-one-subject-out cross-validation is performed to choose the number of selected features by mRMR which is set to 726 for facial movement dynamics and to 377 for head movement dynamics.

For an optimal use of the information given by the facial and head movements, depression severity was assessed by late fusion of separate SVM classifiers. This is done by multiplying the probabilities $s_{i,j}$, output of the SVM for each

TABLE I

CLASSIFICATION ACCURACY (%) - COMPARISON WITH STATE-OF-THE-ART

Method	Modality	Accuracy (%)	Weighted Kappa
J. Cohn *et al.* [7]	Facial movements	59.5	0.43
S. Alghowinem *et al.* [1]	Head movements	53.0	0.42
Dibeklioglu *et al.* [9]	Facial movements	64.98	0.50
Dibeklioglu *et al.* [9]	Head movements	56.06	0.40
Dibeklioglu *et al.* [8]	Facial movements	72.59	0.62
Dibeklioglu *et al.* [8]	Head movements	65.25	0.51
Dibeklioglu *et al.* [8]	Facial/Head movements	**77.77**	**0.71**
Ours	**Facial movements**	**66.19**	**0.60**
Ours	**Head movements**	**61.43**	**0.54**
Ours	**Facial/Head movements**	**70.83**	**0.65**

class j, where $i \in \{1, 2\}$ denotes the modality (*i.e.,* facial and head movements). The class $\mathcal{C}$ of each test sample is determined by

$$\mathcal{C} = \underset{j}{arg\,max} \prod_{i=1}^{2} s_{i,j},\ j = 1, \ldots, n_{\mathcal{C}} \quad , \tag{6}$$

where $n_{\mathcal{C}}$ is the number of classes (*i.e.,* depression severity levels).

V. EVALUATION PROCEDURES

A. Dataset

Fifty-seven depressed participants (34 women, 23 men) were recruited from a clinical trial for treatment of depression. At the time of the study, all met DSM-4 criteria [11] for Major Depressive Disorder (MDD). Data from 49 participants was available for analysis. Participant loss was due to change in original diagnosis, severe suicidal ideation, and methodological reasons (*e.g.,* missing audio or video). Symptom severity was evaluated on up to four occasions at 1, 7, 13, and 21 weeks post diagnosis and intake by four clinical interviewers (the number of interviews per interviewer varied).

Interviews were conducted using the Hamilton Rating Scale for Depression (HRSD) [15]. HRSD is a clinician-rated multiple item questionnaire to measure depression severity and response to treatment. HRSD scores of 15 or higher are generally considered to indicate moderate to severe depression; scores between 8 and 14 indicate mild depression; and scores of 7 or lower indicate remission [13]. Using these cut-off scores, we defined three ordinal depression severity classes: moderate to severe depression, mild depression, and remission (*i.e.,* recovery from depression). The final sample was 126 sessions from 49 participants: 56 moderate to severely depressed, 35 mildly depressed, and 35 remitted (for a more detailed description of the data please see [8]).

B. Results

We seek to discriminate three levels of depression severity from facial and head movement dynamics separately and in combination. To do so, we used leave-One-Subject-Out cross validation scheme. Performance was evaluated using two criterion. One was the mean accuracy over the three levels of severity. The other was weighted kappa [6]. Weighted kappa is the proportion of ordinal agreement above what would be expected to occur by chance [6].

TABLE II

CONFUSION MATRIX

	Remission	Mild	Severe
Remission	**60.0**	31.42	8.57
Mild	20.0	**68.57**	11.42
Severe	1.78	14.28	**83.92**

Consistent with prior work [8], average accuracy was higher for facial movement than for head movement. Facial movement was 66.19%, and head movement was 61.43% (see Table. I). When the two modalities were combined, average accuracy increased to 70.83%.

Misclassification was more common between adjacent categories (e.g., Mild and Remitted) than between distant categories (e.g., Remitted and Severe) (Table. II). Highest accuracy was found for the difference between severe and mild depression (83.92%).

Evaluation of the system components. To evaluate our approach to encoding movement dynamics of face and head movement with alternative representations. For facial movement dynamics, we compared the barycentric representation with a Procrustes representation. Average accuracy using Procrustes was 3% lower than that for barycentric representation (Table. III). For head movements, we compared the Lie algebra representation to a vector representation formed by the yaw, roll, and pitch angles. Accuracy decreased by about 2% in comparison with the proposed approach.

To evaluate whether dimensionality reduction using PCA together with spline interpolation improves accuracy, we compared results with and without PCA and spline interpolation. Omitting PCA and spline interpolation decreased accuracy by about 10%.

To evaluate whether mRMR feature selection and choice of classifier contributed to accuracy, we compared results with and without use of a feature selection step for both Multi-SVM with logistic regression classifiers. When mRMR feature selection was omitted, accuracy decreased by about 8%. Similarly, when logistic regression was used in place of Multi-SVM, accuracy decreased by about 7%. This result

TABLE III

EVALUATION OF THE STEPS TO THE PROPOSED APPROACH

Facial shapes representation	**Accuracy (%)**
Pose normalization (Procrustes)	63.69
Barycentric coordinates	**66.19**

Head pose representation	**Accuracy (%)**
Angles head pose representation	59.05
Lie algebra head pose representation	**61.43**

Impact of spline interpolation	**Accuracy (%)**
Without spline interpolation	60.36
With spline interpolation	**70.83**

Impact of PCA on facial movements	**Accuracy (%)**
Without PCA	56.19
With PCA	**66.19**

Impact of feature selection (mRMR)	**Accuracy (%)**
Without feature selection	62.50
With feature selection	**70.83**

Classifiers	**Accuracy (%)**
Logistic regression	62.02
Multi-class SVM	**70.83**

was unaffected by choice of kernel.

Thus, use of the any of the proposed alternatives would have decreased accuracy relative to the proposed method.

C. Interpretation and Discussion

In this section we evaluate the interpretability of the proposed kinematic features (that is, $K_F(t)$ and $K_H(t)$ defined in Eq. 4 and Eq. 5) for depression severity detection. We compute the l2-norm of velocity and acceleration intensities for the face (*i.e.,* $V_F(t)$ and $A_F(t)$) and head (*i.e.,* $V_H(t)$ and $A_H(t)$) curves for each video. Since each video is analyzed independently, we compute the histograms of the velocity and acceleration intensities over 10 samples (videos) from each level of depression severity. This results in histograms of 50000 velocity and acceleration intensities for each depression level.

Fig. 4 shows the histograms of facial and head velocity (top part) and acceleration (bottom part) intensities. Results for face are presented in the left panel and those for head in the right panel. For face, the level of depression severity is inversely proportional to the velocity and acceleration intensities. Velocity and acceleration both increased as participants improved from severe to mild and then to remitted. This finding is consistent with data and theory in depression.

Head motion, on the other hand, failed to vary systematically with change in depression severity (Fig. 4). This finding was in contrast to previous work. Girard and colleagues [14] found that head movement velocity increased when depression severity decreased. A possible reason for this difference may lie in how head motion was quantified. Girard [14] quantified head movement separately for pitch and yaw; whereas we combined pitch, yaw, and also roll. By combining all three directions of head movement, we may have obscured the relation between head movement and depression severity.

The proposed method detected depression severity with moderate to high accuracy that approaches that of state of the art [8]. Beyond the state of the art, the proposed method yields interpretable findings. The proposed dynamic features strongly mapped onto depression severity. When participants were depressed, their overall facial dynamics were dampened. When depression severity lessened, participants became more expressive. In remission, expressiveness was even higher. These findings are consistent with the observation that psychomotor retardation in depression lessens as severity decreases. Stated otherwise, people more expressive with return to normal mood.

It is possible that future work will enable similar interpretation using deep learning. Efforts toward intepretable artificial intelligence are underway (https://www.darpa.mil/program/explainable-artificial-intelligence). Until that becomes possible, the proposed approach might be considered. Alternatively, it may be most informative to combine approaches such as the one proposed and deep learning.

VI. CONCLUSION AND FUTURE WORK

We proposed a method to measure depression severity from facial and head movement dynamics. Two representations were proposed. An affine-invariant barycentric and Lie algebra representation of facial and head movement dynamics, respectively. The extracted kinematic features revealed strong association between depression severity and dynamics and detected severity status with moderate to strong accuracy.

VII. ACKNOWLEDGEMENTS

We thank J-C. Alvarez Paiva for fruitful discussions on barycentric coordinate representation. Research reported in this publication was supported in part by the U.S. National Institute Of Nursing Research of the National Institutes of Health under Award Number R21NR016510, the U.S. National Institute of Mental Health of the National Institutes of Health under Award Number MH096951, and the U.S. National Science Foundation under award IIS-1721667. The content is solely the responsibility of the authors and does not necessarily represent the official views of the sponsors.

REFERENCES

[1] S. Alghowinem, R. Goecke, M. Wagner, G. Parkerx, and M. Breakspear. Head pose and movement analysis as an indicator of depression. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, pages 283–288, 2013.

[2] A. Beck, C. Ward, M. Mendelson, J. Mock, and J. Erbaugh. An inventory for measuring. *Archives of general psychiatry*, 4:561–571, 1961.

[3] 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.

[4] M. Berger. Geometry, vol. i-ii, 1987.

[5] M. L. Boas. *Mathematical methods in the physical sciences*. Wiley, 2006.

[6] J. Cohen. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213, 1968.

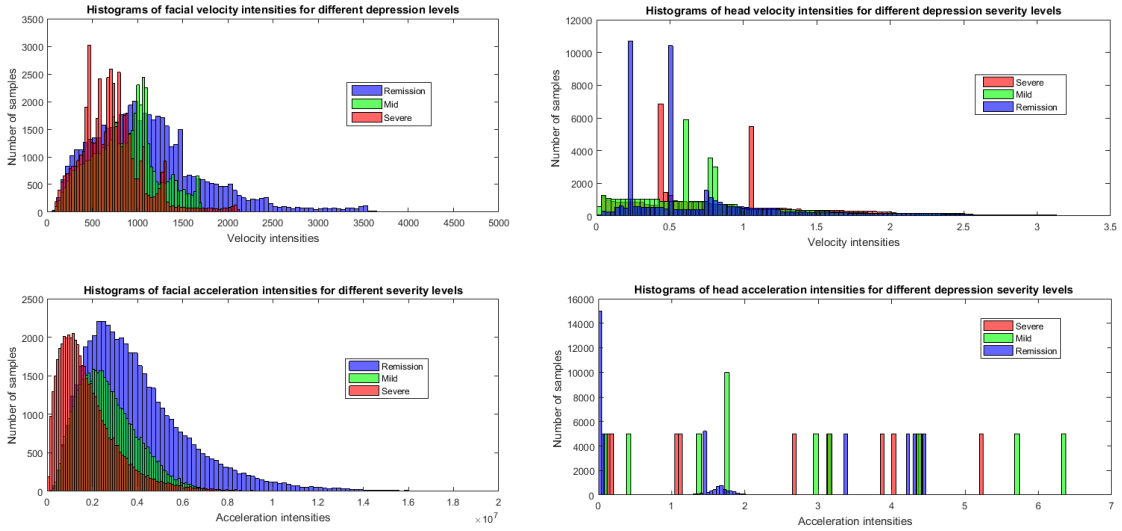

Fig. 4. Histograms of velocity and acceleration intensities for facial (left) and head (right) movements. Psychomotor retardation symptom is well captured by the introduced kinematic features, especially with those computed from the facial movements.

[7] J. F. Cohn, T. S. Kruez, I. Matthews, Y. Yang, M. H. Nguyen, M. T. Padilla, F. Zhou, and F. De la Torre. Detecting depression from facial actions and vocal prosody. In *3rd International Conference on Affective Computing and Intelligent Interaction*, pages 1–7, 2009.

[8] H. Dibeklioglu, Z. Hammal, and J. F. Cohn. Dynamic multimodal measurement of depression severity using deep autoencoding. *IEEE journal of biomedical and health informatics*, 2017.

[9] H. Dibeklioglu, Z. Hammal, Y. Yang, and J. F. Cohn. Multimodal detection of depression in clinical interviews. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, Seattle, WA, USA, November 09 - 13, 2015*, pages 307–310, 2015.

[10] P. Ekman, W. V. Freisen, and S. Ancoli. Facial signs of emotional experience. *Journal of personality and social psychology*, 39(6):1125, 1980.

[11] M. B. First, R. L. Spitzer, M. Gibbon, and J. B. Williams. *Structured clinical interview for DSM-IV axis I disorders - Patient edition (SCID-I/P, Version 2.0)*. Biometrics Research Department, New York State Psychiatric Institute, New York, NY, 1995.

[12] H.-U. Fisch, S. Frey, and H.-P. Hirsbrunner. Analyzing nonverbal behavior in depression. *Journal of abnormal psychology*, 92(3):307, 1983.

[13] J. C. Fournier, R. J. DeRubeis, S. D. Hollon, S. Dimidjian, J. D. Amsterdam, R. C. Shelton, and J. Fawcett. Antidepressant drug effects and depression severity: A patient-level meta-analysis. *Journal of the American Medial Association*, 303(1):47–53, 2010.

[14] J. M. Girard, J. F. Cohn, M. H. Mahoor, S. M. Mavadati, Z. Hammal, and D. P. Rosenwald. Nonverbal social withdrawal in depression: Evidence from manual and automatic analyses. *Image and vision computing*, 32(10):641–647, 2014.

[15] M. Hamilton. A rating scale for depression. *Journal of neurology, neurosurgery, and psychiatry*, 23(1):56–61, 1960.

[16] S. Jayasumana, M. Salzmann, H. Li, and M. Harandi. A framework for shape analysis via hilbert space embedding. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 1249–1256. IEEE, 2013.

[17] L. A. Jeni, J. F. Cohn, and T. Kanade. Dense 3D face alignment from 2D videos for real-time use. *Image and Vision Computing*, 58:13–24, 2017.

[18] J. Joshi, R. Goecke, G. Parker, and M. Breakspear. Can body expressions contribute to automatic depression analysis? In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 1–7, 2013.

[19] A. Kacem, M. Daoudi, and J.-C. Alvarez-Paiva. Barycentric Representation and Metric Learning for Facial Expression Recognition. In *IEEE International Conference on Automatic Face and Gesture Recognition* , Xi'an, China, May 2018.

[20] 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 *International Conference on Computer Vision*, October 2017.

[21] H. Peng, F. Long, and C. Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence*, 27(8):1226–1238, 2005.

[22] F. Perronnin, J. Sánchez, and T. Mensink. Improving the fisher kernel for large-scale image classification. In *European Conference on Computer Vision*, pages 143–156. Springer, 2010.

[23] B. Renneberg, K. Heyn, R. Gebhard, and S. Bachmann. Facial expression of emotions in borderline personality disorder and depression. *Journal of behavior therapy and experimental psychiatry*, 36(3):183–196, 2005.

[24] J. Rottenberg, J. J. Gross, and I. H. Gotlib. Emotion context insensitivity in major depressive disorder. *Journal of abnormal psychology*, 114(4):627, 2005.

[25] J. Saragih and R. Goecke. Iterative error bound minimisation for aam alignment. In *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, volume 2, pages 1196–1195. IEEE, 2006.

[26] G. E. Schwartz, P. L. Fair, P. Salt, M. R. Mandel, and G. L. Klerman. Facial expression and imagery in depression: an electromyographic study. *Psychosomatic medicine*, 1976.

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

[28] M. Valstar, B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, R. Cowie, and M. Pantic. Avec 2014: 3d dimensional affect and depression recognition challenge. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, pages 3–10. ACM, 2014.

[29] R. Vemulapalli, F. Arrate, and R. Chellappa. Human action recognition by representing 3D skeletons as points in a lie group. In *IEEE Conference Computer Vision and Pattern Recognition*, pages 588–595, 2014.

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

[31] J. R. Williamson, T. F. Quatieri, B. S. Helfer, G. Ciccarelli, and D. D. Mehta. Vocal and facial biomarkers of depression based on motor incoordination and timing. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, pages 65–72. ACM, 2014.

[32] Z. Zivkovic. Improved adaptive Gaussian mixture model for background subtraction. In *17th International Conference on Pattern Recognition, ICPR 2004, Cambridge, UK, August 23-26, 2004.*, pages 28–31, 2004.